RICE UNIVERSITY

Bayesian Methods for the Analysis of Microbiome Data

by

William Duncan Wadsworth

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Approved, Thesis Committee:

Michele Guindani
Associate Professor of Statistics
University of California, Irvine

Marina Vannucci, Chair
Noah Harding Professor of Statistics
Rice University

Katherine Bennett Ensor
Professor of Statistics
Rice University

Christopher P. Fagundes
Assistant Professor of Psychology
Rice University

Houston, Texas
October, 2016
ABSTRACT

Bayesian Methods for the Analysis of Microbiome Data

by

William Duncan Wadsworth

Bacteria, archaea, viruses, and fungi are present in large numbers both on and inside of our bodies. On average, only one in ten of “our” cells contain human DNA. The other 90% belong to a tremendous diversity of microbes, some of which are fundamentally related to health and disease mechanisms as documented in numerous recent biomedical studies (Turnbaugh et al., 2009; The Human Microbiome Project, 2012b; Knights et al., 2013; Arpaia et al., 2013; Pickard et al., 2014). Some of these microbes are beneficial while others are detrimental, and, since their abundances are poorly understood, identifying microbes associated with interesting phenotypes is of great importance. However, due to the complexity of these systems and certain characteristics of the data there are still limited numbers of appropriate statistical tools available for such a task.

In this research work I will describe the basic features of microbiome abundance data and present two new modeling approaches that can be used to address some of the challenges presented by this data type. The first approach accomplishes a data integration and model selection goal by associating covariates with microbiome data. The second provides a method of correcting for multiple hypotheses as is common when testing for differential species abundance between experimental or observational conditions. We illustrate the performances of both methods in simulation studies, and
in applications to freely available datasets. Finally, we further discuss their potential
in microbiome research and possible future extensions.
Acknowledgments

Lots of people to thank! To be completed.
Contents

Abstract ii
Acknowledgments iv
List of Illustrations vii
List of Tables x

1 Introduction 1
1.1 Data acquisition ........................................... 2
1.2 Data characteristics ...................................... 4
  1.2.1 Dimensionality and Zero-inflation ......................... 4
  1.2.2 Overdispersion .................................. 7
1.3 Overview of Projects ..................................... 13
  1.3.1 Chapter 2 ........................................ 13
  1.3.2 Chapter 3 ........................................ 14
  1.3.3 Chapter 4 ........................................ 15

2 An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome data 16
2.1 Background .................................................. 16
2.2 Methods ..................................................... 19
  2.2.1 Dirichlet-multinomial regression with variable selection .... 19
  2.2.2 MCMC algorithm .................................. 21
  2.2.3 Comparison study on simulated data ....................... 24
2.2.4 Inferring associations between taxonomic abundances and metabolic pathways .................................................. 26

2.3 Results .................................................................................................................................................. 28
  2.3.1 Simulation study .......................................................................................................................... 28
  2.3.2 Sensitivity analysis ...................................................................................................................... 32
  2.3.3 Data analysis .................................................................................................................................. 33

2.4 Conclusion ............................................................................................................................................. 38

3 Two-groups Poisson-Dirichlet mixtures for multiple testing, with an application to the analysis of microbiome data
  3.1 Introduction ........................................................................................................................................ 41
  3.2 A review of the 2PPD process ......................................................................................................... 45
    3.2.1 Clustering and diversity ................................................................................................................ 48
  3.3 A two-groups 2PPD model ............................................................................................................. 50
  3.4 Posterior inference ........................................................................................................................... 55
  3.5 Identifiability of the two-groups 2PPD model ............................................................................... 59
  3.6 Simulation study ............................................................................................................................... 63
  3.7 Case study: Microbiome data .......................................................................................................... 67
  3.8 Discussion and Conclusion .............................................................................................................. 73
  3.9 Appendix ......................................................................................................................................... 75
    3.9.1 Full conditional ............................................................................................................................ 75
    3.9.2 Proof of Proposition 1 .............................................................................................................. 75
    3.9.3 Sensitivity results ...................................................................................................................... 76

4 Conclusions and discussion .................................................................................................................. 80
Illustrations

1.1 Visualizations of abundance tables with (top) 579 samples and 10,155 OTUs and (bottom) 579 samples and 1,788 OTUs. Zeros are represented as grey squares and Table (top) contains 93.2% while Table (bottom) contains 69.5%. The counts are represented on the log scale for ease of visualization.

1.2 Simulation experiment comparing probability models for the analysis of microbiome data. The dashed line in each subplot shows the 1:1 line.

1.3 Boxplots showing the distribution of E. coli abundance across country of origin and age class. Taxon counts have been square-root transformed to aid in visualization.

2.1 Schematic overview of the proposed integrative Bayesian approach for the application to data from the Human Microbiome Project. The observed data counts (right) are regressed on the available covariates (left), through a variable selection approach, which informs the (unknown) population abundance of each taxon.

2.2 Simulated data: Plot of the marginal posterior probabilities of inclusion, with true associations indicated as red dots.
2.3 Simulated data: Plot of the inferred taxon/covariate association structure based on the median probability model. The magnitude of the association is proportional to the width of the edges and the sign is indicated with a dashed red line for negative and a solid blue line for positive. Circle: taxon; square: covariate.

2.4 Simulated data: Comparison results of selection performances (ROC curves). DMBVS: Dirichlet–Multinomial Bayesian Variable Selection (our method), C&L: Chen and Li, MAPGL: Maximum A Posteriori Bayesian Lasso, CORTEST: Multiplicity Corrected Correlation Tests.

2.5 Simulated data: Comparison results of selection performances.
DMBVS: Dirichlet–Multinomial Bayesian Variable Selection (our method), C&L: Chen and Li, MAPGL: Maximum A Posteriori Bayesian Lasso, CORTEST: Multiplicity Corrected Correlation Tests.

2.6 Real data: Selected negative taxon-by-covariate associations. The magnitude of the association, as captured by the median of the MCMC draws for each $\beta_{pj}$, is proportional to the width of the edges.

2.7 Real data: Selected positive taxon-by-covariate associations. The magnitude of the association, as captured by the median of the MCMC draws for each $\beta_{pj}$, is proportional to the width of the edges.

2.8 Real data: Marginal posterior probabilities of inclusion, with selected associations indicated as red dots.

3.1 Identifiability of the two-groups 2PPD model: Simulation Results under the three Scenarios in Table 3.1, for $\mu_1 \in \{0, 1\}$ and $w \in \{0.67, 1\}$. See Section 3.5 for details.
3.2 Microbiome data case study: Histogram of 564 z-scores obtained from the case term ($\beta_1$) in the Negative Binomial generalized linear mixed effects model with point-wise posterior density estimates for the null and non-null distributions. .......................... 68
3.3 Monte Carlo estimates of the posterior probability of each taxon belonging to the non-null distribution. .......................... 71
Tables

1.1 Example microbiome data for four samples and four microbes at an ambiguous taxonomic level. A typical abundance table would extend to hundreds or thousands of taxa and tens or hundreds of samples.

1.2 Percentages of successfully fit taxa and percentage of minimum AIC between models for each response distribution across each taxon. For 11.2% of taxa, the algorithm failed and no model was able to fit the data.

2.1 Simulated data: Performance assessment for two different scenarios, characterized by different values of the dispersion parameter $\psi$. Values are rounded averages over thirty replicates. Results for accuracy, FPR, FNR, and MCC are based on the median probability model. DMBVS: Dirichlet–Multinomial Bayesian Variable Selection (our method); C&L: Chen and Li; MAPGL: Maximum A Posteriori Bayesian Lasso; CORTEST: Multiplicity Corrected Correlation Tests.

2.2 Simulated data: Sensitivity analysis for varying values of the prior expected value of $p_{pj}$, $m$, and the slab variance $r^2_{pj}$, and for two different scenarios, characterized by different values of the dispersion parameter $\psi$. Values are averages over 30 replicates.
2.3 Real data: Selection results using a BFDR of 0.1. The text in the KEGG column is hyperlinked to the KEGG orthology database for a more complete description. Taxon names start with “g.”, “f.” or “o.” which stand for genus, family, or order, respectively, and correspond to the lowest taxonomic classification available.

3.1 Identifiability of the two-groups 2PPD model: Hyper-parameter values for the $\tilde{p}_j$, $j = 0, 1$ used in the illustration discussed in Section 3.5, with the corresponding prior expected number of clusters when $n = 1,000$ and the values of Simpson’s diversity indexes.

3.2 Simulation study: sensitivity results across different settings for $\sigma_0$ and $\sigma_1$ for the three simulation scenarios considered in Section 3.6 ($w = 0.9$). The values in the table represent the average $F_1$ score, false positive rate (FPR), and true positive rate (TPR) over 30 replicates with corresponding standard deviations.

3.3 Simulation study: comparison between the proposed two-groups 2PPD model and three other multiple comparison methods for the three simulation scenarios considered in Section 3.6 ($w = 0.9$). The values in the table represent the average $F_1$ score over 30 replicates with corresponding standard deviations.

3.4 Microbiome data case study: differentially abundant taxa with negative $z$-scores indicating less abundance in the children with moderate to severe diarrhea. Most are well known commensal bacteria, e.g. *Prevotella* spp. and *Clostridium* spp.

3.5 Microbiome data case study: differentially abundant taxa with positive $z$-scores indicating greater abundance in the children with moderate to severe diarrhea. Most are well known pathogenic bacteria, e.g. *Shigella* spp. and *E. coli.*
3.6 Simulation study: sensitivity results across different settings for $\sigma_0$ and $\sigma_1$ for the three simulation scenarios considered in Section 3.6 ($w = 0.95$). The values in the table represent the average $F_1$ score, false positive rate (FPR), and true positive rate (TPR) over 30 replicates with corresponding standard deviations.

4.1 List of possible interactions between microbiome and host. Note that each taxon will interact with every other taxon in one of these ways too.
Chapter 1

Introduction

The microbiome is the meta-community of bacteria, viruses, and fungi that live on and inside of the human body. These communities are large; some estimates count as many as ten bacterial cells in the gut for every human one (Guarner and Malagelada, 2003). Because each species has a distinct genome this represents a vastly larger quantity of genetic information than does the human genome alone. Additionally, there are numerous sites including the skin, lungs, vagina, and gut, that carry very distinct communities. The compositions of these microbial communities are complex and there is substantial evidence to suggest that their status has an impact on human health and disease. The presence or absence of certain bacteria have been associated with phenotypes such as diabetes, obesity, eczema, irritable bowel syndrome, and pneumonia (Kassinen et al., 2007; Turnbaugh et al., 2009; Giongo et al., 2011; Bousbia et al., 2012; Kong et al., 2012; Arpaia et al., 2013; Pickard et al., 2014). Furthermore, these communities adapt to their surroundings (Flores et al., 2014) and this potential for change may allow them to be manipulated for therapeutic purposes (Sonnenburg and Fischbach, 2011; Haiser and Turnbaugh, 2012). Evidence for the manipulability of the microbiome, however, remains elusive Dror et al. (2015).

These factors highlight the promise and the difficulty of studying the microbiome. In this chapter, we will describe microbiome data, how it is obtained, some of its unique characteristics, and some commonly used methods for its analysis. In presenting these methods we will highlight the strengths and weaknesses of each and
motivate the methods proposed in the following chapters.

1.1 Data acquisition

Historically, information about the composition of microbial communities was obtained by culturing bacteria on media known to be favored by certain species. Only species that could be cultured in the lab – and species whose metabolic preferences were known – could be measured. More recently, the ability to use modern high-throughput sequencing technology has allowed researchers to employ targeted gene sequencing to profile the microbial population in a sample (DeSantis et al., 2003). This profiling makes use of so-called marker genes to single out specific microbes and estimate their absolute abundance. Marker genes such as the 16S rDNA gene in bacteria are highly conserved within microbial groups so all individuals in that group carry a copy. However, much like a genetic fingerprint, each species has subunit sequences within the marker gene that are distinct and indicative of that species. By reading and enumerating all the distinct subunit sequences an estimate of the number of distinct taxa may be made. In other words, if \( n \) of a particular sequence of nucleotides is observed in a sample, it may be assumed that \( n \) individuals from a certain taxon are present in that sample.

The process of enumerating and counting the transcripts requires extensive use of dedicated bioinformatics pipelines (Hamady and Knight, 2009). Because sequences do not always cleave at the same location they must be clustered by similarity into Operational Taxonomic Units (OTUs) rather than species. These clusters are commonly defined at 97% similarity, though more stringent classifications, e.g. 99%, are sometimes used. The mapping of these nucleotide sequence clusters to biologically recognized taxa is done using bioinformatics software such as QIIME (Caporaso et al.,
2010) or mothur (Schloss et al., 2009) that make use of curated online databases such as SILVA (Pruesse et al., 2007) and Greengenes (DeSantis et al., 2006). Because the sequences are mapped to these clusters the concept of a species is not always well defined and we will often refer to biological classifications ambiguously as taxa. Furthermore, since there is a natural hierarchy in the way that biological organisms are classified, i.e. species, genus, family, etc., given a table of OTU abundances one may “glom” them into any one of these classifications and perform the analysis at that level. Doing so reduces the dimensionality of the data and simplifies interpretation since relatively little is known about the vast majority of OTUs. Note also, that OTUs are commonly considered to be the finest grain of classification possible with 16S rDNA sequencing and fall below the level of genus.

<table>
<thead>
<tr>
<th></th>
<th>Taxon 1</th>
<th>Taxon 2</th>
<th>Taxon 3</th>
<th>Taxon 4</th>
<th>⋯</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>⋯</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>⋯</td>
</tr>
<tr>
<td>Sample 3</td>
<td>54</td>
<td>0</td>
<td>1</td>
<td>53</td>
<td>⋯</td>
</tr>
<tr>
<td>Sample 4</td>
<td>59</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>⋯</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
</tbody>
</table>

Table 1.1 : Example microbiome data for four samples and four microbes at an ambiguous taxonomic level. A typical abundance table would extend to hundreds or thousands of taxa and tens or hundreds of samples.

There are numerous biological and computational details to profiling microbial populations; for an overview of the process of obtaining abundance data from extracted DNA, see Morgan and Huttenhower (2012) and (Thomas et al., 2012). Nu-
merous challenges are also present. The entire process of conducting a microbiome experiment includes sample acquisition and storage, DNA extraction, PCR amplification, clustering, and the mapping in order to obtain a final count of OTUs. The potential for introduction of biases is present throughout this process (Brooks et al., 2015) and processing of microbiome data is a very active area of research.

In this work we focus on statistical methods for analyzing one of the final product of these processes: the abundance table. An example abundance table is shown in Table 1.1 where columns represent species of microbes and rows represent samples from different patients. By remaining agnostic to how samples are procured and processed we ensure that the modeling approaches may still be applied as the technology rapidly progresses.

1.2 Data characteristics

Microbiome data has several qualities that complicate analyses using standard parametric statistical tools. Dimensionality, count-based measurement, zero-inflation, and overdispersion all characterize the microbiome due to the species-sampling nature of counting many taxa across individuals. The non-uniform distribution of individuals within species has been studied since at least 1943 when Fisher et al. published on the species distribution of butterflies (Fisher et al., 1943). Here we describe these characteristics and use them to justify the modeling approaches presented in later chapters.

1.2.1 Dimensionality and Zero-inflation

Diversity, as measured by richness (absolute number of taxa) and evenness (uniformity of taxa abundance), varies widely across the sites of the human body (Spor
et al., 2011). For example, the most diverse site is the gut, where as many as 1,000 species may be found in a healthy individual (Sears, 2005), while the healthy vaginal microbiome is characterized primarily by twenty species in a single genus *Lactobacillus* (Huang et al., 2014). In the modeling context, richness in a sample corresponds to dimensionality since the abundance of each taxon is represented as a column in the abundance table.

Zero-inflated data exhibits more observed mass at zero than would be expected from standard probability models. Zero-inflation in microbiome data can occur because sequencing sometimes fails to identify rare species and instead records a zero. This biases the estimates since it is unknown whether a zero count represents the absence of a species in a sample or just a counting error for that species. Additionally, as more and more samples are included in a particular study, the greater the likelihood of counting a rare species (Good, 1953). This effect increases the dimensionality and the absolute number of zeros in an abundance table. In an extreme case single patient may record a rare species but the rest of that species’ column must be necessarily filled with zeros. Taxa inclusion thresholds for handling such sampling issues are discussed below.

Figure 1.1 highlights the dimensionality and zero-inflation by visualizing an abundance table from a study of temporal variability across multiple sites (Flores et al., 2014). Note that this is an extreme case since different sites, here represented in the same abundance table, are known to contain vastly different communities (The Human Microbiome Project, 2012b). Figure 1.1 (top) shows an unfiltered abundance table where all sequenced and mapped OTUs are presented, including 2,335 taxa that are present in just one sample. This table contains 93.2% zeros. Since it is likely that
Figure 1.1: Visualizations of abundance tables with (top) 579 samples and 10,155 OTUs and (bottom) 579 samples and 1,788 OTUs. Zeros are represented as grey squares and Table (top) contains 93.2% while Table (bottom) contains 69.5%. The counts are represented on the log scale for ease of visualization.

The rarest taxa are chimera or other sequencing artifacts it is common practice to set a presence threshold for the inclusion of taxa. A common threshold requires that taxa be present in some percentage of samples in order to be considered for analysis. Figure 1.1 (bottom) shows the same abundance table with such an exclusion threshold applied. Only taxa present in 10% or more of the samples are shown resulting
in a table with 1,788 OTUs and 69.5% zeros. Despite the thresholding, the data is clearly still high-dimensional and zero-inflated.

### 1.2.2 Overdispersion

Overdispersion, also common in microbiome data, is a property of data wherein the mean is greater than the variance. Single-parameter family exponential family distributions impose the constraint that the variance be some function of the mean. Thus, single-parameter exponential family distributions cannot adequately model overdispersed data without using hierarchical structure or variance stabilizing transformations (Anscombe, 1948). To demonstrate the overdispersion in microbiome data and useful probability models for applications we conduct a simulation study again using the data from Flores et al. (2014). Figure 1.2 shows results from an experiment in the style of “posterior predictive checking” (Gelman et al., 2013, p. 143). This procedure involves a) estimating parameters from specific models, b) simulating data from those models and estimated parameters, and c) visually or statistically comparing the simulated results to the original dataset.

For each of the taxa $y_j$, $j = 1, \ldots, J$, the standard estimates of the mean $\bar{y}_j$ and the variance $s_j^2$ were obtained and plotted on the log scale, see Figure 1.2 (top left). A standard linear regression of the means and variances is shown as a red line but is displayed in other colors for the various probability models. For the univariate Poisson($\lambda$) and Negative Binomial($\theta$, $r$) models the parameters $\lambda_j$ and ($\theta_j$, $r_j$) were estimated, respectively, for each of the taxa using maximum likelihood. For the multivariate Multinomial($\mu_1, \ldots, \mu_J, n$) and Dirichlet-Multinomial($\gamma_1, \ldots, \gamma_J, n$), the parameters ($\mu_1, \ldots, \mu_J$) and ($\gamma_1, \ldots, \gamma_J$) were also estimated using maximum likelihood. Using the parameter estimates from each probability model, data was simulated for each of
the taxa \( y_j^r, j = 1, \ldots, J \) with the same number of observations as the original data \((n = 579)\) and, as before, the standard estimates of the mean \( \bar{y}_j^r \) and the variance \( s_{j}^2 \) were obtained and plotted on the log scale.

Figure 1.2: Simulation experiment comparing probability models for the analysis of microbiome data. The dashed line in each subplot shows the 1:1 line.

For the Poisson model, \( \mathbb{E}(y_j) = \mathbb{V}(y_j) = \lambda_j \), which is clear in Figure 1.2 (top middle) since the estimated variances lie along the 1:1 line. For the Negative Binomial model, \( \mathbb{E}(y_j) = r_j \theta_j \), but \( \mathbb{V}(y_j) = r_j (1 - \theta_j)/\theta_j^2 \). We can make the multiplicative overdispersion factor clear by expressing \( \mathbb{V}(y_j) = \mathbb{E}(y_j)[(1 - \theta_j)/\theta_j^3] \). Thus, the Negative Binomial model allows for the overdispersion graphically represented in Figure 1.2.
(top right). Similarly, for each taxa the expected values are $E(y_j) = n\mu_j$ and $E(y_j) = n\gamma_j$ for the Multinomial and Dirichlet-Multinomial distributions, respectively. However, both variances accommodate overdispersion with $V(y_j) = E(y_j)(1 - \mu_j)$ for the Multinomial and $V(y_j) = E(y_j)(1 - \gamma_j)(n + \gamma_+)/(1 + \gamma_+)$ where $\gamma_+ = \sum_{j=1}^{J} \gamma_j$ for the Dirichlet-Multinomial. This can be seen in Figure 1.2 (bottom left and bottom middle) where the simulated data $y^*_j$ clearly displays overdispersion. In Figure 1.2 (bottom right) the simulated data has been removed and only the fitted linear regressions are plotted showing how well the Negative Binomial and Dirichlet-Multinomial distributions mimic the overdispersion pattern of microbiome data.

Further evidence for the suitability of the Negative Binomial distribution can be demonstrated by fitting a more complicated model. Using data from Pop et al. (2014) we here explore the use of marginal models and the suitability of zero-inflated models for microbiome data. Like the approach in the previous demonstration, a marginal model is one that is independently fit to each of the taxa $j = 1, \ldots, J$. The data from Pop et al. (2014) are a case-controlled study of post-diarrheal disruption in children from several low-income countries. Stool samples were obtained from 992 children between the ages of 0 and 59 months, 508 of whom had recently suffered from moderate to severe diarrhea, while the remaining 484 children acted as age-matched controls. The samples were obtained in Mali, The Gambia, Kenya, and Bangladesh and case/control proportions were approximately equal from each country. This data is also used in Chapter 3 to demonstrate a multiple comparison correction method. Using the full design of the study allows us to fit a marginal Negative Binomial generalized linear mixed effects model (GLMM) on the taxonomic abundances $y_{ij}$, where $j = 1, \ldots, J$, indexes the taxa, and $i = 1, \ldots, n$ indexes the observations. To account for heterogeneity of absolute species abundance between patients, we let $s_i$
denote the total species counted in sample \(i\). We also let \(x_{ij}\) denote a binary covariate such that \(x_{ij} = 1\) if observation \(y_{ij}\) is from one of the cases and \(x_{ij} = 0\) otherwise.

We consider the following model:

\[
y_{ij} \sim NB(\mu_{ij}, \alpha_j), \quad j = 1, \ldots, J; \quad i = 1, \ldots, n,
\]

\[
\log(\mu_{ij}) = s_i + \beta_{0j} + \beta_{1j} x_{ij} + u_{\text{country},j} + u_{\text{case},j} + u_{\text{age class},j},
\] (1.1)

where \(\alpha_j\) represents a taxon-specific dispersion parameter, \(\beta_{0j}\) represents a taxon-specific effect capturing the abundance of the taxon in the control group. To address group-specific variability, \(u_{\text{country},j}\), \(u_{\text{case},j}\), and \(u_{\text{age class},j}\) are independent normally distributed zero-mean random effects with variances \(\sigma^2_{\text{country},j}\), \(\sigma^2_{\text{case},j}\), and \(\sigma^2_{\text{age class},j}\).

A visualization of the design is shown in Figure 1.3 for the species \(E. \ coli\). In this case, a clear and systematic difference in abundance of \(E. \ coli\) can be seen across case status and age class. The increased abundance of \(E. \ coli\) in the affected children is predictable since some strains are well-known pathogens responsible for severe gastrointestinal disturbance. The clear downward trend in abundance by age class contradicts earlier hypotheses explained by a decrease in the frequency of breastfeeding and an increase in exposure to contaminated food and water as a child ages (Ochoa et al., 2009).

The generalized linear mixed-effects model in Equation (1.1) was fit to each of 660 species using the R package \texttt{glmmADMB} (Skaug et al., 2016). \texttt{glmmADMB} fits the GLMM by maximizing Laplace approximations to the marginal likelihood (Fournier et al., 2012). For each of the species, the distribution of the response \(y_{ij}\) was assumed to come from each of the following: Poisson, Zero-inflated Poisson, Negative Binomial, Zero-inflated Negative Binomial, Type 1 Negative Binomial, and Type 1 Zero-inflated Negative Binomial. The Type 1 Negative Binomial is an alternate parametrization of
Figure 1.3: Boxplots showing the distribution of E. coli abundance across country of origin and age class. Taxon counts have been square-root transformed to aid in visualization.

the Negative Binomial that in some cases eases parameter estimation (Cameron and Trivedi, 2013). In presenting the following results, we appeal to the so-called “folk theorem of statistical computing”. This “theorem” states that if software fails to fit a certain model the problem is most likely with the model itself.

In 74 species the software failed to fit any of the above listed models while for the remaining 586 species the algorithm obtained convergence for at least one of the response distributions. The first column of Table 1.2 gives the percentage, out of 660...
species, of models that converged for each response distribution. For the remaining models the algorithm failed to converge after an extraordinary number of iterations or failed outright for numerical reasons. For each fitted model the Akaike Information Criteria (AIC) was calculated and compared within the individual taxa. The second column of Table 1.2 shows the percentage of taxa for which each model had the minimum AIC among the other successfully fitted models. Despite the large quantity of zeros in the data the evidence for using zero-inflated models is weak. For nearly half of the taxa the standard Negative Binomial distribution had the minimum AIC and of the remaining taxa 3/5 were best fit by the Type 1 Negative Binomial distribution. The pattern for successfully estimating the taxon-wise parameters was similar with the Negative Binomial fitting nearly 3/4 of the taxa and the Type 1 Negative Binomial fitting just under 2/3.

<table>
<thead>
<tr>
<th>Model</th>
<th>% successful</th>
<th>% minimum AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Zero-inflated Poisson</td>
<td>10.1</td>
<td>01.0</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>71.3</td>
<td>49.1</td>
</tr>
<tr>
<td>Zero-inflated Negative Binomial</td>
<td>50.7</td>
<td>18.0</td>
</tr>
<tr>
<td>Negative Binomial, Type 1</td>
<td>61.5</td>
<td>31.5</td>
</tr>
<tr>
<td>Zero-inflated Negative Binomial, Type 1</td>
<td>1.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 1.2: Percentages of successfully fit taxa and percentage of minimum AIC between models for each response distribution across each taxon. For 11.2% of taxa, the algorithm failed and no model was able to fit the data.
Clearly these results only give an informal demonstration of the suitability of certain models for microbiome data. However taken as a whole these results encourage the use of certain response distributions over others. In the following projects we propose the use of the Negative Binomial and Dirichlet-Multinomial distributions, along with hierarchical structure through appropriate prior distributions, to account for the discussed features of microbiome data.

1.3 Overview of Projects

The projects in the following chapters present novel approaches to analyzing microbiome abundance tables in a Bayesian framework. A brief summary of the research follows.

1.3.1 Chapter 2

In Chapter 2, we present an integrative Bayesian approach to variable selection with the Dirichlet-Multinomial regression model. Integrative approaches which aim at associating the composition of the human microbiome with other available information, such as clinical covariates and environmental predictors, are paramount to develop a more complete understanding of the role of microbiome in disease development. By fixing spike-and-slab priors on the regression coefficients we obtain concise summaries of association between an abundance table and a set of associated covariates. The approach allows straightforward incorporation of the covariates through a log-linear regression parametrization of the parameters of the Dirichlet-Multinomial likelihood. Inference is conducted through a Markov Chain Monte Carlo algorithm, and selection of the significant covariates is based upon the assessment of posterior probabilities of inclusions and the thresholding of the Bayesian false discovery rate. We design a
simulation study to evaluate the performance of the proposed method, and then apply our model on a publicly available dataset obtained from the Human Microbiome Project which associates taxa abundances with KEGG orthology pathways. This method compares favorably in simulations to several recently proposed approaches for similarly structured data, in terms of increased accuracy and reduced false positive as well as false negative rates. In the application to the data from the Human Microbiome Project, a close evaluation of the biological significance of our findings confirms existing associations in the literature.

The material presented in Chapter 2 corresponds to a manuscript which has been re-submitted to the journal BMC Bioinformatics. The work was co-authored with Raffaele Argiento (CNR-IMATI, Milan), Michele Guindani (University of California, Irvine), Jessica Galloway-Pena (MD Anderson Cancer Center), Samuel A. Shelbourne (MD Anderson Cancer Center) and Marina Vannucci (Rice University).

1.3.2 Chapter 3

Chapter 3 presents a general Bayesian non-parametric approach to multiple comparison corrections. The simultaneous testing of multiple hypotheses is common in the analysis of high-dimensional datasets such as microbiome abundance tables. The two-groups model, first proposed by Efron (2004), identifies significant comparisons on the basis of the estimation of an empirical null distribution. In the Bayesian nonparametric literature, many approaches have suggested the use of mixtures of Dirichlet Process for the two-groups model. Here, instead we investigate the use of a mixture of two-parameter Poisson Dirichlet Processes (2PPD) as a more flexible way to estimate the distributions in the two-groups model. By making appropriate choices for the hyper-parameters of the processes we can encourage behavior relevant
to the two-groups problem improve the identifiability of the mixture. By means of a simulation study, we compare the performances of our method to those of other recently proposed methods for large-scale inference in the two-groups paradigm. We illustrate the use of our method on a case-control study of the microbial composition of the gastrointestinal tracts in children from poor countries who have been diagnosed with moderate to severe diarrhea.

The material presented in Chapter 3 corresponds to a manuscript that is in preparation and was co-authored with Michele Guindani (University of California, Irvine), Fabrizio Leisen (University of Kent) and Antonio Lijoi (University of Pavia).

1.3.3 Chapter 4

In Chapter 4 we conclude by discussing potential extensions of the proposed methods and areas of future interest for the Bayesian analysis of microbiome abundance data. In particular, due to the dynamic nature of the microbiome, there is a need for dynamic models that can account for longitudinal sampling. In this chapter we will introduce an alternative to count-based analysis approaches. Compositional data analysis (Aitchison, 1982) transforms count-based and compositional data into a variety of simpler mathematical spaces which will be briefly discussed.
Chapter 2

An Integrative Bayesian Dirichlet-Multinomial Regression Model for the Analysis of Taxonomic Abundances in Microbiome data

2.1 Background

The human microbiome is defined as the collection of microorganisms, including bacteria, viruses, and some unicellular eukaryotes, that live in and on our bodies (Morgan and Huttenhower, 2012). Research on the microbiome has grown exponentially in the past few years and it has been argued that the microbiota can be regarded as a “second genome” (Zhu et al., 2010; Grice and Segre, 2012). Indeed, just the human gut microbiota is estimated to be composed of approximately 10^{14} bacterial cells, i.e. ten times more than the total number of human cells in the body (Fraher et al., 2012). The contribution of the human microbiome on several health outcomes has been frequently reported in the literature. For example, microbial dysbiosis in the gut has been linked to irritable bowel syndrome and Crohn’s disease (Abraham and Cho, 2009), type 2 diabetes (Qin et al., 2012), cardiovascular disease (Koeth et al., 2013), and psychological conditions via the so-called “gut-brain axis” (Cryan and O’Mahony, 2011). The composition of microbiota at other body sites have also been associated with conditions such as eczema (Kong et al., 2012) and pre-term labor (Romero et al., 2014). This stream of research holds great potential for a better understanding of many mechanistic processes in the development of human diseases,
especially with respect to immune regulation and barrier defense (Devaraj et al., 2013; Ash and Mueller, 2016).

Many of the existing tools for microbial community analysis (e.g., the QIIME platform, Caporaso et al., 2010) bypass those characteristics and rely on nonparametric tests to compare species across different conditions (Wu et al., 2011; Youmans et al., 2015). Other approaches use ordination, e.g. multidimensional scaling, to summarize abundances, and are sometimes employed to link the microbiome data with available clinical covariates and phylogenetic information (Hamady et al., 2010; Fukuyama et al., 2012). In those approaches, the choice of the distance metric is often crucial. The interpretation of biological phenomena can also be challenging in low dimensional projections. Most importantly, distance-based methods do not explicitly quantify the relative importance of significant associations between taxa and covariates, and therefore are of limited use for clinical decisions.

In this manuscript, we consider an integrative Bayesian approach based on the use of Dirichlet-Multinomial (DM) distributions Mosimann (1962) for studying the association between taxa abundance data and available measurements on clinical, genetic and environmental covariates. Recently, La Rosa et al. la Rosa et al. (2012) proposed the use of a DM model for hypothesis testing and power calculations in microbiome experiments. Holmes et. al Holmes et al. (2012) used a finite mixture of DM distributions to directly model the taxa counts, but without incorporating predictors to study the influence of external factors on the microbiome’s abundance. A penalized likelihood approach based on a DM regression model has been proposed instead by (Chen and Li, 2013) to determine significant associations between the microbiome composition and a set of covariates which describe the individual dietary nutrients’ intakes. Similarly, (Chen et al., 2013) develop a structure constrained version of
sparse canonical correlation analysis that integrates compositionalized microbiome data, phylogenetic information, and nutrient information, whereas (Lin et al., 2014) propose penalized regression models to associate compositionalized microbiome data with some phenotype of interest, e.g. body mass index, as a response. However, the use of a constrained optimization approach does not allow to fully characterize the uncertainty in the selection of the significant associations, which is of particular importance, especially when dealing with high-dimensional and highly-correlated data.

Here, we propose a probabilistic modeling approach which both flexibly takes into account the typical features of microbiome count data and also allows for straightforward incorporation of available covariate information within a DM log-linear regression framework. By imposing sparsity inducing spike-and-slab priors on the regression coefficients, our model obtains a parsimonious summary of the effects of the associations and also allows an assessment of the uncertainty of the selection process. We evaluate the performance of our model first on simulated data, where we provide comparisons with methods developed for microbiome or similar type of data. We also illustrate our method on data obtained from the Human Microbiome Project (The Human Microbiome Project, 2012b), to investigate the association between taxonomic abundances and metabolic pathways inferred from whole genome shotgun sequencing reads. It is known that the combination of environmental and host genetic factors shape the composition of the gut microbiota, and these interactions appear to have a significant effects on several biological mechanisms, which may be related, for example, to the individual immunity and barrier defense, as well as metabolism and diet (Benson et al., 2010; Goodrich et al., 2016).
2.2 Methods

We describe our Bayesian variable selection approach for the analysis of microbiome data and their association with a set of available covariates in the context of DM log-linear regression models.

2.2.1 Dirichlet-multinomial regression with variable selection

Let \( y_i = (y_{i1}, \ldots, y_{iJ}) \) indicate the vector of counts representing the taxonomic abundance table obtained from the \( i \)th patient, with \( y_{ij} \) denoting the frequency of the \( j \)th microbial taxon, for \( j = 1, \ldots, J \) and \( i = 1, \ldots, n \). Furthermore, let \( X = (x_1, \ldots, x_P) \) indicate a \( n \times P \) matrix of measurements on \( P \) covariates.

We start by modeling the taxonomic count data with a Multinomial distribution

\[
y_i \mid \phi_i \sim \text{Multinomial}(y_{i+}, \phi_i),
\]  

with \( y_{i+} = \sum_{j=1}^{J} y_{ij} \) the summation of all counts in the vector, and where the parameters \( \phi \)'s are defined on the \( J \) dimensional simplex

\[
S^{J-1} = \{(\phi_1, \ldots, \phi_J) : \phi_j \geq 0, \forall j, \sum_{j=1}^{J} \phi_j = 1\}.
\]

We further impose a conjugate Dirichlet prior on \( \phi \), that is

\[
\phi \sim \text{Dirichlet}(\gamma),
\]

where \( \gamma = (\gamma_1, \ldots, \gamma_J) \) indicates a \( J \)-dimensional vector of strictly positive parameters. An advantage of our hierarchical formulation is that conjugacy can be exploited to integrate \( \phi \) out, obtaining the Dirichlet–Multinomial model, \( y_i \sim \text{DM}(\gamma) \), with probability mass function
\[
f(y_i \mid \gamma) = \left( \frac{y_{i+}}{y_i} \right) \frac{\Gamma(y_{i+} + 1) \Gamma(\gamma_+)}{\Gamma(y_i + \gamma_+)} \prod_{j=1}^{J} \frac{\Gamma(y_{ij} + \gamma_j)}{\Gamma(y_{ij} + 1)},
\]

where \( \gamma_+ = \sum_{j}^{J} \gamma_j \). First described in Mosimann (1962) as the compound multinomial, the DM(\( \gamma \)) allows more flexibility than the Multinomial when encountering overdispersion in multivariate count data, as it induces an increase in the variance by a factor of \((y_+ + \gamma_+)/(1 + \gamma_+)\).

Next, we incorporate the covariates into the modeling via a log-linear regression framework where the DM parameters depend on the available covariates \( X \)'s. More specifically, we define \( \zeta_j = \log(\gamma_j) \) and assume

\[
\zeta_j = \alpha_j + \sum_{p=1}^{P} \beta_{pj} x_p,
\]

\( i = 1, \ldots, n; \ j = 1, \ldots, J. \)

In this formulation, the intercept term \( \alpha_j \) corresponds to the log baseline parameter for the taxon \( j \), whereas the regression parameter \( \beta_{pj} \) captures the effect of the \( p \)th covariate on the abundance for that taxon.

Identifying the significant associations between taxa and covariates in model (2.1)-(2.2) is equivalent to determining the non-zero \( \beta_{pj} \) parameters. One way to address this issue is through variable selection and the use of spike-and-slab mixture priors (George and McCulloch, 1997; Brown et al., 1998). First, we introduce latent binary indicator vectors \( \xi_j = (\xi_{1j}, \xi_{2j}, \ldots, \xi_{pj}) \), such that \( \xi_{pj} = 1 \) if the \( p \)th covariate influences the abundance of the \( j \)th taxon and \( \xi_{pj} = 0 \) otherwise. Then, we write the prior on the \( \beta_{pj} \)'s as

\[
\beta_{pj} \sim \xi_{pj} \mathcal{N}(0, \tau_{pj}^2) + (1 - \xi_{pj}) \delta_0(\beta_{pj}),
\]

(2.3)
where $\delta_0$ denotes a Dirac-delta at 0 and $r_j^2$ is some suitably large value (Smith and Kohn, 1996; Chipman et al., 2001). Large values for $r_j^2$ suggest a flat prior distribution on the location of $\{\beta_{pj} \mid \xi_{pj} = 1\}$ and therefore encourage the selection of relatively large effects. We place Bernoulli priors on the selection indicators $\xi_{pj}$, that is

$$
\pi(\xi_j \mid \mathbf{p}_j) = \prod_{p=1}^{P} p_{pj}^{\xi_{pj}} (1 - p_{pj})^{1-\xi_{pj}}. \tag{2.4}
$$

We also specify Beta hyperpriors on the hyperparameters $p_{pj}$, i.e., $p_{pj} \sim \text{Beta}(a, b)$, as this has been shown to provide an automatic adjustment for multiplicity (Scott and Berger, 2010). This is equivalent to placing a Beta mixed Binomial distribution on $\xi_{pj}$,

$$
\pi(\xi_{pj}) = \int \pi(\xi_{pj} \mid \lambda) \pi(\lambda) d\lambda,
$$

with $\lambda = (a, b)$. In practice, a relatively weakly specification can be obtained by having a “flat” Beta distribution with the sum of its two parameters being two. Thus we set $a$ and $b$ so that the prior expected mean value is $m = a/(a + b)$ with $a + b = 2$. Finally, we assume normal priors on the $\alpha_j$’s, i.e. $\alpha_j \sim \mathcal{N}(0, s_j^2)$ with large values for $s_j^2$ encoding a diffuse prior. Figure (2.1) provides an overview of the proposed integrative modeling approach, with reference to the application on data from the Human Microbiome Project described later.

### 2.2.2 MCMC algorithm

We implement a stochastic search Markov Chain Monte Carlo (MCMC) algorithm for posterior inference that employs a Gibbs scan to sample the non-zero regression coefficients (Savitsky et al., 2011). We encourage an efficient sampling by employing
Figure 2.1: Schematic overview of the proposed integrative Bayesian approach for the application to data from the Human Microbiome Project. The observed data counts (right) are regressed on the available covariates (left), through a variable selection approach, which informs the (unknown) population abundance of each taxon.

a component-wise adaptive Metropolis algorithm (Roberts and Rosenthal, 2009) as described below. A generic iteration of the MCMC algorithm comprises the following steps:

1. **Update of $\alpha$:** This is a Metropolis-Hastings step with a symmetric random walk proposal $\alpha_j' \sim \mathcal{N}(\alpha_j, t_\alpha^2)$, for $j = 1, \ldots, J$.

2. **Joint update of $(\xi, \beta)$:** We sample these parameters jointly via a Gibbs scan that employs a Metropolis acceptance step. For each $j = 1, \ldots, J$ and $p = 1, \ldots, P$:

   - if $\xi_{pj} = 1$: propose $\xi'_{pj} = 0$ and $\beta_{pj}' = 0$.
   - if $\xi_{pj} = 0$: propose $\xi'_{pj} = 1$ and then propose $\beta_{pj}'$ following an adaptive Metropolis-Hastings scheme.
\[
\beta'_{pj} \sim 0.95 \mathcal{N}(\beta_{pj}, 2.38^2 \times \hat{\sigma}^2_{pj} / J \times P) + 0.05 \mathcal{N}(\beta_{pj}, 0.01 / J \times P)
\]

where \(\hat{\sigma}^2_{pj}\) is the current estimate of the variance of the target distribution.

The value of \(\hat{\sigma}^2_{pj}\) is updated using a recursive formula as in (Haario et al., 2005) on all the previous draws for \(\beta_{pj}\).

- Accept \((\xi'_{pj}, \beta'_{pj})\) with probability

\[
a = \min \left\{ \frac{\pi(\xi'_{pj}, \beta'_{pj} \mid \xi_{pj}, \beta_{pj}, \text{else})}{\pi(\xi_{pj}, \beta_{pj} \mid \xi_{pj}, \beta_{pj}, \text{else})} \right\},
\]

where \(\xi_{pj} = (\xi'_{1,j}, \ldots, \xi'_{p-1,j}, \xi_{p+1,j}, \ldots, \xi_{pj})\), and

\(\beta'_{pj} = (\beta'_{1,j}, \ldots, \beta'_{p-1,j}, \beta_{p+1,j}, \ldots, \beta_{pj})\).

For posterior inference, we are interested in identifying the relevant associations between taxa and covariates as captured by the selection indicators \(\xi_{pj}\)'s and the corresponding regression coefficients \(\beta_{pj}\)'s. Estimates of the marginal posterior probabilities of inclusion (PPIs) of the latent indicators \(\xi_{pj}\) can be calculated by counting the number of times that each taxon/covariate association is included across the MCMC iterations. A selection of the significant associations can then be made by choosing those elements that have marginal PPIs greater than a specific value, for example greater than 0.5 for the median probability model of Barbieri and Berger (2004). Another choice for the threshold which controls for multiplicity (Newton et al., 2004) relies on an estimated pre-specified Bayesian false discovery rate \(\alpha\) calculated as

\[
\widehat{\text{FDR}}(c) = \frac{\sum_{p=1}^{P} \sum_{j=1}^{J} (1 - \widehat{\text{PPI}}_{pj}) D_{pj}}{\sum_{p=1}^{P} \sum_{j=1}^{J} D_{pj}},
\]
where $D_{pj} = \mathbb{1}(\mathbb{P}(\pi_{pj} > c))$. An optimal threshold $c'$ can be found for error rate $\alpha$ by choosing $c'$ such that $\hat{\text{FDR}}(c') < \alpha$. Estimates of the non-zero regression coefficients $\beta_{pj}$ can also be calculated by averaging over the sampled MCMC values.

In order to compare selection performance of different methods, we calculate accuracy, false positive rate (FPR), false negative rate (FNR) and Matthews correlation coefficient (MCC), across 30 replicated datasets. We define accuracy as $\text{ACC} = (TP + TN)/(P + N)$, with $TP$ the number of true positives out of $P$ selected and $TN$ the number of true negatives out of $N$ not selected. The false negative rate is calculated as $\text{FNR} = FN/(FN + TP)$, the false positive rate as $\text{FPR} = FP/(FP + TN)$, and the Matthews correlation coefficient as

$$\text{MCC} = \frac{TP/N - S \times P}{\sqrt{TPS(1 - S)(1 - P)}},$$

with $N = TN + TP + FN + FP$, $P = \frac{TP + FP}{N}$ and $S = \frac{TP + FN}{N}$ Matthews (1975).

We further computed receiving operating curves (ROC) to compare the performance of the selection procedure across the different methods.

### 2.2.3 Comparison study on simulated data

We carry out a simulation study to assess the performance of our model and compare results to alternative methods. We compare performance with three approaches to variable selection for multivariate count-response regression: the penalized approach of Chen and Li (2013), the factorized maximum a posteriori (MAP) Gamma Lasso of Taddy (2013) and the false discovery rate-corrected full pair-wise correlation test of Spearman’s $\rho$ as done in Wu et al. (2011). When fitting the Bayesian Gamma Lasso method of Taddy (2013), model selection was done using the
minimum AIC, while for Chen and Li’s approach the minimum BIC was calculated with the group penalty set to 20%. We also fit the method of Chen and Li to the untransformed data. The false discovery rate threshold for the Spearman’s correlation tests was set to 0.05.

In simulating data, we set \( n = 100, P = 50 \) and \( J = 50 \), and chose \( P_r = 9 \) and \( J_r = 5 \) to obtain a total number of relevant taxon/covariate associations equal to 25. We simulated the covariate matrix \( \mathbf{X} \) according to a Multivariate-Normal(0, \( \Sigma \)) with \( \Sigma_{i,j} = \rho^{|i-j|} \) and \( \rho = 0.4 \). We then drew each vector \( \mathbf{y}_i \) of counts from a Multinomial(\( N, \pi \)) distribution with row sum \( N \in [1,000,2,000] \) and with parameters \( \pi = (\pi_1, \ldots, \pi_J) \) set to \( \pi_j = (\gamma_j/\gamma_+) \times (1 - \psi)/\psi \), with overdispersion parameter \( \psi = 0.01 \) and with \( \gamma_j = \exp\{\alpha_j + \mathbf{X}\beta_j\} \) and \( \gamma_+ = \sum_{j=1}^{J} \gamma_j \). We sampled the non-zero \( \beta_{pj} \)'s from the intervals \( \pm[0.5,1.0] \) and the intercept parameters from a Uniform(-2.3, 2.3). Below we report performance results as averages over 30 replicated simulated datasets.

When running the MCMC, we used a vague prior for the intercept by setting the variance parameter to \( s^2_{\pi_j} = 10 \). Similarly, we set \( r^2_{\pi_j} = 10 \), to provide sufficiently vague prior information on the non-zero log-linear regression coefficients. Finally, we set \( m = 0.01 \) (or \( a = 0.02 \) and \( b = 1.98 \)), resulting in a sparse prior mean on selected associations of 1% of the total.

We provide comments on the sensitivity of the selection results to the choice of these hyperparameters in the Section below. We ran the MCMC algorithm for 10,000 iterations and thinned to every fifth iteration. On a single dataset, the C code took approximately 31.5 minutes to run on an Intel Xeon E5-2630 2.30GHz processor. We assessed convergence visually and via the Geweke diagnostic (Geweke, 1992) as implemented in the R package \texttt{coda}. Convergence was checked for a) the number of
active variables in each iteration and b) the samples from each of the selected $\beta_{pj}$.
The five number summary of the 25 Geweke $z$-scores was (-3.43, -1.06, -0.63, 0.71, 1.98).

2.2.4 Inferring associations between taxonomic abundances and metabolic pathways

We demonstrate our approach on publicly available data obtained from the Human Microbiome Project (HMP) website The Human Microbiome Project (2012b) from which we use 79 samples from healthy individuals. The $Y$ matrix in our model contains 16S rRNA microbial counts from stool samples at the genus taxonomic level. As common in microbiome studies, the genera abundances ($Bacteroides$, $Prevotella$, etc.) were filtered by requiring each genus to be present in at least 5% of the samples. This procedure removes extremely low-abundance genera leaving 80 genera for the analysis. From the same 79 individuals, we obtained KEGG orthology group abundances which are used as the matrix of covariates $X$ of our model. The KEGG orthology groups were reconstructed from metagenomic shotgun sequencing (WGS) using the HMP Unified Metabolic Analysis Network (HUMAnN) pipeline (Abubucker et al., 2012) and were also provided on the HMP website. These values represent inferred abundances of biochemical functional groups and metabolic pathways present due to the shotgun sequenced reads of bacterial and non-bacterial genes in the sample. To reduce correlation among the covariates we used average linkage clustering on the correlation matrix of the KEGG groups and chose one representative from each cluster, according to its relevance to microbiome research, leaving 76 columns in $X$. Finally, the columns in $X$ were mean centered and scaled to unit variance. Though the HMP sampled 300 individuals for several time points and over many sites, there were
relatively few samples that included the WGS used to obtain the KEGG orthology data. Thus, when joining the samples from the 16S rRNA data and the KEGG orthology data, a total of 79 matched samples remained.

We used the same hyperparameter settings as in the simulation study, that is $s^2_{pj} = 10$ and $r^2_{pj} = 10$ and set $m = 0.01$, resulting in a sparse mean selection prior of 1% of the total 6,080 possible associations. The MCMC algorithm described in Section 2.2.2 above was run for 500,000 iterations and thinned to every 100th draw. We assessed convergence visually and via the Geweke diagnostic (Geweke, 1992) as implemented in the R package `coda`. The five number summary of the Geweke $z$-scores for the 26 $\beta_{pj}$’s was (-3.83, -1.19, 0.15, 1.46, 3.38).

![Figure 2.2: Simulated data: Plot of the marginal posterior probabilities of inclusion, with true associations indicated as red dots.](image)
Figure 2.3: Simulated data: Plot of the inferred taxon/covariate association structure based on the median probability model. The magnitude of the association is proportional to the width of the edges and the sign is indicated with a dashed red line for negative and a solid blue line for positive. Circle: taxon; square: covariate.

2.3 Results

2.3.1 Simulation study

In Figure 2.2 we show the plot of the marginal PPIs of the $P \times J$ elements $\zeta_{pj}$, obtained by computing the proportion of times that $\zeta_{pj} = 1$ across all iterations, after burn-in. The selected median model, corresponding to a threshold of 0.5 on the PPIs, is depicted in Figure 2.3. In this Figure, the magnitude of the association, as captured by the estimated $\beta_{pj}$’s, is proportional to the width of the edges, with dashed red lines indicating negative associations and solid blue lines positive associations. This model fails to select a true association between variable 4 and taxon 4 and falsely includes an association between variable 17 and taxon 4, resulting in a false positive rate of
0.0004 and a false negative rate of 0.04. The value of the AUC for this replicate was 0.99.

Figure 2.4 : Simulated data: Comparison results of selection performances (ROC curves). DMBVS: Dirichlet-Multinomial Bayesian Variable Selection (our method), C&L: Chen and Li, MAPGL: Maximum A Posteriori Bayesian Lasso, CORTEST: Multiplicity Corrected Correlation Tests.

Figure 2.4 illustrates the selection performance of the proposed method, by plotting the average ROC curves over the 30 replicated datasets \( \psi = 0.01 \) for each of the methods included in the comparison. The Figure shows that our proposed model outperforms the competing methods in terms of achieved average true and false positive rates. Boxplots of the selection performance values over the 30 replicated datasets with \( \psi = 0.01 \) are reported in Figure 2.5 and show that our proposed model either
outperforms or is commensurate with the competing methods.

![Graph showing comparison results of selection performances]

Figure 2.5: Simulated data: Comparison results of selection performances. DMBVS: Dirichlet–Multinomial Bayesian Variable Selection (our method), C&L: Chen and Li, MAPGL: Maximum A Posteriori Bayesian Lasso, CORTEST: Multiplicity Corrected Correlation Tests.

As an additional comparison, together with the total number of correctly identified regression parameters, which we term “overall recovery”, we also looked at the “taxon-wise recovery”, which we defined as the correct recovery of any element from one of the $J$ taxa. Thus, recovery for overall selection occurs for $P \times J$ elements while taxon-wise selection occurs for $J$ elements. Table 2.1 reports average values for
Table 2.1: Simulated data: Performance assessment for two different scenarios, characterized by different values of the dispersion parameter \( \psi \). Values are rounded averages over thirty replicates. Results for accuracy, FPR, FNR, and MCC are based on the median probability model. DMBVS: Dirichlet–Multinomial Bayesian Variable Selection (our method); C&L: Chen and Li; MAPGL: Maximum A Posteriori Bayesian Lasso; CORTEST: Multiplicity Corrected Correlation Tests.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DMBVS</td>
<td>MAPBL</td>
<td>C&amp;L</td>
<td>CORTEST</td>
<td>DMBVS</td>
<td>MAPBL</td>
<td>C&amp;L</td>
<td>CORTEST</td>
</tr>
<tr>
<td>( \psi = 0.01 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.96</td>
<td>0.91</td>
<td>0.85</td>
<td>0.95</td>
</tr>
<tr>
<td>FNR</td>
<td>0.05</td>
<td>0.10</td>
<td>0.12</td>
<td>0.31</td>
<td>0.00</td>
<td>0.46</td>
<td>0.43</td>
<td>0.02</td>
</tr>
<tr>
<td>FPR</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>0.01</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>MCC</td>
<td>0.93</td>
<td>0.64</td>
<td>0.67</td>
<td>0.73</td>
<td>0.89</td>
<td>0.66</td>
<td>0.50</td>
<td>0.85</td>
</tr>
<tr>
<td>( \psi = 0.1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.92</td>
<td>0.86</td>
<td>0.81</td>
<td>0.92</td>
</tr>
<tr>
<td>FNR</td>
<td>0.39</td>
<td>0.58</td>
<td>0.28</td>
<td>0.63</td>
<td>0.24</td>
<td>0.73</td>
<td>0.52</td>
<td>0.37</td>
</tr>
<tr>
<td>FPR</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>0.01</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>MCC</td>
<td>0.72</td>
<td>0.42</td>
<td>0.54</td>
<td>0.56</td>
<td>0.73</td>
<td>0.40</td>
<td>0.38</td>
<td>0.70</td>
</tr>
</tbody>
</table>

accuracy, FPR, FNR and MCC, averaged across the 30 replicated datasets, for both overall and taxon-wise recovery. These results show that our method in particular outperforms competing methods for taxon-wise recovery. In the same Table we report results for a more challenging simulated scenario, obtained with a higher value of
the overdispersion parameter ($\psi = 0.1$). As expected, the increase in overdispersion makes the selection task more difficult for all methods. However, our method still outperforms or is commensurate with the competing methods, even in the presence of considerable overdispersion.

### 2.3.2 Sensitivity analysis

Since our proposal requires the choice of a number of hyperparameters, it is important to investigate how sensitive the results are to varying parameter sets. Therefore, we conclude our simulation study by briefly discussing the sensitivity of the results to the prior specifications. In general, we found that results were robust to the prior choices on the intercept parameters, $\alpha_j$, while, as expected, some sensitivity was observed with respect to the variance hyperparameters of the *spike-and-slab* prior (2.3) on the regression coefficients, $\beta_{pj}$, and the hyperparameters of the Beta priors on $p_{pj}$. In Table 2.2 we report results obtained by considering a full grid of values for the prior expected value of $p_{pj} m \in \{0.005, 0.01, 0.05\}$ and the slab variance $r^2_{pj} \in \{1, 10, 100\}$. In the Additional File, we further report the ROC curves corresponding to each choice of the parameters. With only 25 truly non-zero $\beta_{pj}$'s, out of 2,500 parameters, small increases in false positive rates can drastically decrease the Matthews correlation coefficient. Thus imposing some sparsity by using a smaller value for $m$ improves overall performance while larger values of $m$ allow for more false positives. Similarly, when the slab variance is small, e.g. $r^2_{pj} = 1$, there is more prior density close to zero which allows small but insignificant variables to be selected. Conversely, when the slab variance is large, e.g. $r^2_{pj} = 100$, false positives are less likely but false negatives increase since the prior density is more evenly spread over the support.
Table 2.2: Simulated data: Sensitivity analysis for varying values of the prior expected value of $p_{pj}$, $m$, and the slab variance $r_{pj}^2$, and for two different scenarios, characterized by different values of the dispersion parameter $\psi$. Values are averages over 30 replicates.

<table>
<thead>
<tr>
<th></th>
<th>$m = 0.005$</th>
<th>$m = 0.01$</th>
<th>$m = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{pj}^2$</td>
<td>$r_{pj}^2$</td>
<td>$r_{pj}^2$</td>
<td>$r_{pj}^2$</td>
</tr>
<tr>
<td></td>
<td>$= 1$</td>
<td>$= 10$</td>
<td>$= 100$</td>
</tr>
<tr>
<td>$\psi = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>FPR</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>FNR</td>
<td>0.02</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>MCC</td>
<td>0.69</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>AUC</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\psi = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>FPR</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>FNR</td>
<td>0.26</td>
<td>0.37</td>
<td>0.46</td>
</tr>
<tr>
<td>MCC</td>
<td>0.53</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>AUC</td>
<td>0.96</td>
<td>0.96</td>
<td>0.93</td>
</tr>
</tbody>
</table>

2.3.3 Data analysis

Figure 2.8 shows the traceplot of the number of included taxon/covariate associations and the plot of the marginal PPIs of the $P \times J$ elements $\xi_{pj}$, obtained by computing the proportion of times that $\xi_{pj} = 1$ across all iterations, after burn-in. Here the median model, corresponding to a threshold of 0.5 on the PPIs, selects 92 associations. Among those, 26 have a marginal PPI greater than 0.98, which corresponds to a
Figure 2.6: Real data: Selected negative taxon-by-covariate associations. The magnitude of the association, as captured by the median of the MCMC draws for each \( \beta_{pj} \), is proportional to the width of the edges.

Bayesian FDR of 0.1. These 26 associations are listed in Table 2.3, together with the corresponding estimated regression coefficients, and depicted in Figures 2.7 and 2.6, for positive and negative associations, respectively. In these Figures, the magnitude
Figure 2.7: Real data: Selected positive taxon-by-covariate associations. The magnitude of the association, as captured by the median of the MCMC draws for each $\beta_{pj}$, is proportional to the width of the edges.
positive associations.

The text in the KEGG column of Table 2.3 is hyperlinked to the KEGG orthology database (Kanehisa et al., 2016), for more complete descriptions of the selected pathways. A close investigation of the biological significance of the associations identified by our model reveals several interesting characteristics and affirms the relevance of these associations. Commensal microbiota that inhabit the human gut are proficient at scavenging glycans and polysaccharides, including those in plants, such as starches or cellulose, animal-derived tissues (glycosaminoglycans and N-linked glycans), and glycans from host mucus (O-linked glycans) (Koropatkin et al., 2012). *Ruminococcus* spp. are known to participate in both resistant starch and glycosaminoglycan degradation (Koropatkin et al., 2012; Walker et al., 2011). It has been reported that long-term consumption of diets rich in protein and animal fat were associated with an enterotype primarily containing increased *Bacteroides* and *Ruminococcus* species.
Additionally, *Ruminococccus torques* and *Ruminococcus gnavus* have been shown to degrade mucins (Crost et al., 2013). Thus, it is logical that *Ruminococcus*, which is one of the noteworthy genera involved in glycosaminoglycan degradation, would be negatively associated to glycosaminoglycan biosynthesis (ko00053) (Table 2.3). Similarly, *Parabacteroides* which is also negatively associated with N-glycan biosynthesis (ko00513), is involved in deglycosylation and utilization of N-glycans (Cao et al., 2014). Also, among the associations identified for the glycan pathways, *Prevotella* was negatively associated with mucin type O-glycan biosynthesis (ko00512). In the literature, *Prevotella* has implications for mucosal homeostasis, as some *Prevotella* spp. express a unique mucin-desulfating glycosidase that can hydrolyze GlcNAc residues on mucin-type O-glycans, and thus is important for mucin degradation (Rho et al., 2005). Other associations affirmed through the literature included that of *Bacteroides* with naphthalene degradation (ko00626). It has been reported that *Bacteroidetes* possess the capability to degrade polycyclic aromatic hydrocarbons such as naphthalene (Hilyard et al., 2008). Associations of *Ruminococcus* with pyruvate metabolism (ko00620) are also supported, as phosphoenolpyruvate carboxykinase was previously reported to be associated with *Ruminococcus flavefaciens* in the rumen (Schöcke and Weimer, 1997). Another supported association is that of *Prevotellaceae* with sulfur metabolism. L-cysteine desulphhydrase enzymes have been characterized in *Prevotella intermedia* (Yano et al., 2009). Additionally, glycosulfatase enzymes have been described in *Prevotella* (Wright et al., 2000). Equally interesting is the selection of pathways that are expected to be omnipresent among many bacteria, such as glycolysis/gluconeogenesis (ko00010), as glycolysis occurs, with variations, in nearly all organisms, both aerobic and anaerobic. Thus, it is not surprising that taxa like *Bacteroides* and *Clostridiales* are associated with glycoly-
sis/gluconeogenesis as they are abundant taxa within the gut microbiome. A number of selected associations exhibited negative coefficients, see for example Bacteriodes with glycolysis/gluconeogenesis.

Given the complexity of metabolic pathways and the process of mapping specific genes to pathways, some of the selected associations are unexpected, and might be due to the 16S abundances that were made available at the HMP site and the mapping of metagenomic sequences to specific KEGG orthology groups by HUMAnN. For example, several species of Ruminococcus are known to participate in butanoate (butyrate) metabolism (Takahashi et al., 2016), Dialister spp. have phenylalanine arylamidase activities (Jumas-Bilak et al., 2005), and Prevotella spp. are known to participate in pyruvate metabolism (Takahashi and Yamada, 2000; Ruan et al., 2015). Since those associations should be driven exclusively by bacterial genes, it is interesting that we find significant associations between the abundance of certain bacterial taxa and KEGG pathways that are primarily reported among eukaryotic species (i.e., T-cell receptor signaling, hedgehog signaling, pathways in cancer, etc.). Indeed, although precautionary steps are performed, the HMP consortium reported that human contaminants are found in 50%-90% of the sequences (The Human Microbiome Project, 2012a). These unexpected findings suggest the need for further investigations and validation.

2.4 Conclusion

Herein, we have developed a Bayesian approach to the Dirichlet-Multinomial regression models that allows for the selection of significant associations between covariates and taxa from a microbiome abundance table by imposing spike-and-slab priors on the log-linear regression coefficients of the model. We have applied our model to simulated
data and compared performances with methods developed for similar applications.

We have illustrated the performance of our method using publicly available data on
taxonomic abundances and metabolic pathways inferred from whole genome shotgun
sequencing reads, which we obtained from the Human Microbiome Project website.
Our results have revealed interesting links between specific taxa (i.e. genera) and
particular metabolic pathways, which we have validated via existing literature.

Several extensions of our model are possible. Because some habitats, e.g. the
gut, are thought to have highly variable dynamics, longitudinal sampling may be pre-
ferred to cross-sectional sampling since it may give a better sense of long-term trends
(Faith et al., 2013). Thus, incorporating repeated samples with specified correlation
structures in the linear predictor could produce additional insights. Another impor-
tant aspect of microbiome data, which is receiving attention from researchers, is the
heterogeneity in community structure across samples, as this can be an indication
of the existence of “enterotypes” (Koren et al., 2013; Wang et al., 2014). This can
be addressed within our modeling framework by employing Bayesian nonparamet-
ric models that would allow to cluster selected associations across partitions of the
samples. These extensions are currently under investigation.
<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Pathway</th>
<th>taxon</th>
<th>MPPI</th>
<th>$\beta_{pj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ko04660</td>
<td>T cell receptor signaling pathway</td>
<td>g.Subdoligranulum</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>ko04512</td>
<td>ECM-receptor interaction</td>
<td>g.Subdoligranulum</td>
<td>1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>ko00680</td>
<td>Methane metabolism</td>
<td>g.Sutterella</td>
<td>1.00</td>
<td>-1.47</td>
</tr>
<tr>
<td>ko05200</td>
<td>Pathways in cancer</td>
<td>o.Bacteroidales</td>
<td>1.00</td>
<td>-1.04</td>
</tr>
<tr>
<td>ko04540</td>
<td>Gap junction</td>
<td>o.Bacteroidales</td>
<td>1.00</td>
<td>-0.92</td>
</tr>
<tr>
<td>ko00534</td>
<td>Glycosaminoglycan biosynthesis</td>
<td>g.Ruminococcus</td>
<td>1.00</td>
<td>-0.56</td>
</tr>
<tr>
<td>ko00053</td>
<td>Ascorbate and aldarate metabolism</td>
<td>g.Ruminococcus</td>
<td>1.00</td>
<td>-0.46</td>
</tr>
<tr>
<td>ko00650</td>
<td>Butanoate metabolism</td>
<td>g.Ruminococcus</td>
<td>1.00</td>
<td>-0.55</td>
</tr>
<tr>
<td>ko00513</td>
<td>Various types of N-glycan biosynthesis</td>
<td>g.Parabacteroides</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>ko00500</td>
<td>Starch and sucrose metabolism</td>
<td>g.Parabacteroides</td>
<td>1.00</td>
<td>-0.61</td>
</tr>
<tr>
<td>ko00904</td>
<td>Diterpenoid biosynthesis</td>
<td>g.Dialister</td>
<td>1.00</td>
<td>-1.21</td>
</tr>
<tr>
<td>ko00360</td>
<td>Phenylalanine metabolism</td>
<td>g.Dialister</td>
<td>1.00</td>
<td>-2.18</td>
</tr>
<tr>
<td>ko00626</td>
<td>Naphthalene degradation</td>
<td>g.Bacteroides</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>ko0010</td>
<td>Glycolysis / Gluconeogenesis</td>
<td>g.Bacteroides</td>
<td>1.00</td>
<td>-0.57</td>
</tr>
<tr>
<td>ko00513</td>
<td>Various types of N-glycan biosynthesis</td>
<td>g.Prevotella</td>
<td>1.00</td>
<td>0.77</td>
</tr>
<tr>
<td>ko00512</td>
<td>Mucin type O-Glycan biosynthesis</td>
<td>g.Prevotella</td>
<td>1.00</td>
<td>-1.06</td>
</tr>
<tr>
<td>ko00620</td>
<td>Pyruvate metabolism</td>
<td>g.Prevotella</td>
<td>1.00</td>
<td>-1.76</td>
</tr>
<tr>
<td>ko04340</td>
<td>Hedgehog signaling pathway</td>
<td>g.Ruminococcus</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>ko04660</td>
<td>T cell receptor signaling pathway</td>
<td>g.Roseburia</td>
<td>1.00</td>
<td>0.46</td>
</tr>
<tr>
<td>ko00983</td>
<td>Drug metabolism</td>
<td>g.Sutterella</td>
<td>1.00</td>
<td>-1.20</td>
</tr>
<tr>
<td>ko0260</td>
<td>Glycine, serine and threonine metabolism</td>
<td>o.Bacteroidales</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>ko00330</td>
<td>Arginine and proline metabolism</td>
<td>g.Ruminococcus</td>
<td>0.99</td>
<td>0.39</td>
</tr>
<tr>
<td>ko01057</td>
<td>Biosynthesis of type II polyketide products</td>
<td>o.Bacteroidales</td>
<td>0.99</td>
<td>0.76</td>
</tr>
<tr>
<td>ko00620</td>
<td>Pyruvate metabolism</td>
<td>g.Ruminococcus</td>
<td>0.99</td>
<td>0.56</td>
</tr>
<tr>
<td>ko00920</td>
<td>Sulfur metabolism</td>
<td>f.Prevotellaceae</td>
<td>0.99</td>
<td>1.36</td>
</tr>
<tr>
<td>ko00010</td>
<td>Glycolysis / Gluconeogenesis</td>
<td>o.Clostridiales</td>
<td>0.98</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 2.3 : Real data: Selection results using a BFDR of 0.1. The text in the KEGG column is hyperlinked to the KEGG orthology database for a more complete description. taxon names start with “g.”, “f.” or “o.” which stand for genus, family, or order, respectively, and correspond to the lowest taxonomic classification available.
Chapter 3

Two-groups Poisson-Dirichlet mixtures for multiple testing, with an application to the analysis of microbiome data

3.1 Introduction

The recent availability of high-dimensional data in domains as diverse as genomics, imaging, and astronomy, has brought to the forefront of scientific investigation the necessity of finding powerful methods for screening a large number of measurements at once, in order to validate pre-assumed hypotheses. In statistics, this type of large-scale inference has been typically associated with the so-called multiple testing problem (Efron, 2010), that is the need for controlling the type I error rate over all the hypotheses being tested. As a simple example, a naive application of a sequence of two sample t-tests at the $\alpha = 0.05$ significance level over multiple independent comparisons will induce a probability of at least one false discovery as high as 0.4 for just 10 comparisons. Therefore, significance level adjustments such as the Bonferroni correction and the false discovery rate control approaches of Benjamini and Hochberg (1995) and Storey (2002) are now widely used by scientists. Although the methods developed to address such problems can be applied to any multiple testing problem, a great part of the recent development has been motivated by high-throughput genomic technologies such as mRNA microarrays, RNA-seq, and more recently, 16S rRNA microbial community analysis. In 16S rRNA differential abundance studies
the goal is to determine systematic changes in abundance across many taxa between two conditions. For example, in our application in Section 3.7, we consider a microbiome study comparing data from healthy children and children recently affected by a gastrointestinal disruption.

Here, we focus on the two-groups model for large scale hypothesis testing, which has been proposed by Efron (2004). See also Efron (2008). For illustration, assume that the observations are suitably defined difference scores \( z_i, i = 1, \ldots, n \) corresponding to a large number of distinct hypotheses. The two-groups model assumes that the difference scores come from either a null \( (f_0) \) or a non-null \( (f_1) \) distribution, i.e. each score is drawn from a mixture,

\[
z_i \sim w f_0 + (1 - w) f_1,
\]

for some weight \( w \in (0, 1) \), and some probability (density) functions \( f_0 \) and \( f_1 \). While the null component is typically assumed to be a standard normal distribution, in reality the true null distribution may depart from that assumption. Therefore, Efron proposes the estimation of an “empirical null” distribution to adequately capture the range of parameter values coherent with the null hypothesis. Each decision about any tested hypotheses is then taken with respect to the empirical distribution, in lieu of the theoretical standard normal distribution.

Bayesian nonparametric methods have focused on the two-group model as a general reference for the multiple testing problem due to their ability to provide flexible estimates of the unknown null and alternative distributions. The Dirichlet process has often been used either to estimate the densities of \( f_0 \), or \( f_1 \) or both, as well as to cluster the difference scores into common "expression" levels. For example, Dahl and Newton (2007) have proposed a methodology that uses the Dirichlet process to cluster gene expressions in order to increase testing sensitivity. This work was
extended by Kim et al. (2009) who apply a spike-and-slab base distribution to the Dirichlet process prior. This effectively clusters genes into a null point-mass at zero or a non-null Gaussian which allows for inference on “sharp hypotheses” such as a gene’s null/non-null status and allows for the estimation of the posterior probabilities of such hypotheses. Other approaches estimate both $f_0$ and $f_1$ from the data. For example, Do et al. (2005) apply Dirichlet process priors to $f_0$ and $f_1$ and model the $z$-scores in a hierarchical fashion, where the null distribution is characterized by a unimodal Gaussian base measure, whereas the alternative distribution considers a two–component Gaussian mixture base measure. Similarly, Tang et al. (2007) propose a Dirichlet process mixture model on the distribution of $P$-values, assuming a null uniform distribution. More recently, Kottas and Fellingham (2010) consider mixtures of location-shifted symmetric distributions, where parametric priors are used for the location parameters and for the mixture proportion, whereas a common nonparametric scale uniform Dirichlet process mixture prior is used to estimate the random symmetric density. For large scale hypothesis testing, Martin and Tokdar (2012) also use a flexible hierarchical nonparametric approach, where the null distribution is assigned a Normal distribution with unknown mean and variance, and the non-null distribution is assigned a location mixture of normals. Scott (2009) proposes a fully Bayesian nonparametric approach where both $f_0$ and $f_1$ are assigned Dirichlet process priors for conducting multiple testing of autoregressive time series. A recently proposed approach by Zhao et al. (2014) further improves on the two-groups model by using a mixture Dirichlet process prior on both the null and non-null components, but letting the group membership indicator be a weighted Ising prior in order to explore dependencies in gene sub-networks. One of the appealing features of the two-groups model is that the resulting inference is immediately amenable to
interpretation in a decision theoretic framework. For example, Efron (2004) describes a local version of the false discovery rate (local fdr), which represents the posterior probability that a difference score \( z_i \) is generated according to the null hypothesis, defined as \( fdr(z) = w f_0(z) / f_1(z) \). Then, the selection of interesting scores is conducted by flagging all \( z_i \)'s such that \( fdr(z_i) < \alpha, \alpha \in (0, 1) \), which allows control of the Benjamini–Hochberg FDR at level \( \alpha \). Alternatively, the decision problem could be formulated by optimizing suitable loss functions which define a weighted compound of expected false positive and false negative decisions. In most cases, the optimal decision rule is then obtained by thresholding the posterior probability of the alternative hypothesis, or a transformation thereof [e.g., see Müller et al. (2006) and Guindani et al. (2009)].

In this manuscript, we investigate the use of a mixture prior of two-parameter Poisson–Dirichlet (2PPD) processes as a flexible class of prior processes which can be used to estimate the unknown distributions in the two-groups model. The 2PPD process, introduced by Pitman (1996) and also known as the Pitman-Yor process in the machine learning literature, is a generalization of the Dirichlet process, and is characterized by two parameters: a concentration parameter \( \theta \in \mathbb{R}^+ \), as in the Dirichlet Process, and also a parameter \( \sigma \in [0, 1) \). In particular, we discuss how large values of \( \sigma \) imply that the estimates of the unknown distributions may not depart substantially from their prior expectations, and therefore we suggest the use of 2PPD processes characterized by high values of \( \sigma \) to model the empirical null distribution \( f_0 \). On the contrary, small values of \( \sigma \) allow the 2PPD process to vary more widely, and therefore they are more appropriate to elicit the uncertainty related to the non-null distribution. We derive the expression of the exchangeable partition probability function (EPPF) of the mixture of 2PPD processes, which describes the distribution
of the exchangeable random partition induced by the 2PPD process. In particular, we show that conditional on the assignment of the observations to either the null or the alternative hypothesis, the respective random partitions are independent. In Section 3.4, we show that this facilitates posterior inference through the use of MCMC algorithms, which take into account the conditional independence of the partition. We further illustrate how the choice of the parameters of the 2PPD process might facilitate the estimation of the allocation probabilities and the identifiability of the mixture components, a well-known problem when using mixture models for describing population heterogeneity (Teicher, 1963; Titterington et al., 1985). We discuss the performance of our method with respect to existing state-of-the-art approaches for large-scale multiple comparison problems by means of a simulation study and then illustrate the use of the proposed mixture of 2PPD process on publicly available data from a microbiome study, where the aim was to identify differences in the microbial composition of the gastrointestinal tracts of children from poor countries, who have been diagnosed with moderate to severe diarrhea.

### 3.2 A review of the 2PPD process

In this Section, we provide an overview of the 2PPD process, with particular regard to its clustering properties. Let $X_1, \ldots, X_n$ denote a sample of $n$ observations, which are drawn from a sequence of exchangeable random elements $X_1, X_2, \ldots$, taking value in a complete and separable metric space $\mathcal{X}$ endowed with its Borel $\sigma$-algebra $\mathcal{F}$. This is a common assumption in Bayesian modeling, which allows characterization of the distribution of the sequence of observations, by virtue of the de Finetti representation
theorem, as follows:

\[ X_i \mid \tilde{p} \sim \tilde{p} \quad i = 1, \ldots, n, \]

\[ \tilde{p} \sim Q, \]

for any \( n \geq 1 \), and for \( \tilde{p} \) a random probability measure, with distribution \( Q \) defined on the space \( \mathcal{P}(\mathbb{X}) \) of probability measures on \( \mathbb{X} \). In a Bayesian setup \( Q \) represents the prior distribution and the model is said to be parametric whenever \( Q \) degenerates on a finite dimensional subspace of \( \mathcal{P}(\mathbb{X}) \); otherwise, the model is denoted as nonparametric.

In order to introduce the notation we use in later Sections, we define the 2PPD by characterizing \( Q \) as a completely random measure (CRM). A CRM \( \mu \) is a random element taking values in the space of boundedly finite measures on \( \mathbb{X} \), such that for any two disjoint sets \( A \) and \( B \) in \( \mathcal{X} \) the random variables \( \mu(A) \) and \( \mu(B) \) are independent. A CRM is characterized by the so-called Lévy-Khintchine representation, which provides an expression for the Laplace functional transform of \( \mu \), i.e.

\[ \mathbb{E}[e^{-t\mu(A)}] = \exp \left\{ -\int_{\mathbb{R}^+ \times A} [1 - e^{-ts}] \nu(ds, dx) \right\}, \quad (3.2) \]

for any measurable function \( f : \mathbb{X} \to \mathbb{R} \), such that \( \int |f|d\mu < \infty \), and where \( \nu(ds, dx) \) denotes the Lévy intensity of \( \mu \), that is a measure on \( \mathbb{R}^+ \times \mathbb{X} \), such that \( \nu(ds, dx) = \alpha(dx) \rho(s \mid x)ds \), with \( \int_{\mathbb{X}} \rho(s \mid x)\alpha(dx) < \infty \), for any \( s > 0 \), and \( \int_{\mathbb{R}^+} \min(s, 1) \rho(s \mid x)ds < \infty \), for any \( x \) in \( \mathbb{X} \). If \( \rho(s \mid x) = \rho(s) \), for any \((s, x) \in \mathbb{R}^+ \times \mathbb{X} \), then \( \mu \) is said to be a homogeneous CRM. Homogeneous CRMs are widely employed in Bayesian nonparametrics to define a variety of random probability measures. For example, the commonly used Dirichlet Process is obtained by considering the normalized completely random measure \( \mu/\mu(\mathbb{X}) \), after setting \( \rho(s) = s^{-1}e^{-s} \), i.e. as a normalized Gamma CRM, since \( \mu(A) \sim \text{Gamma}(1, \alpha(A)) \). Alternatively, by setting
\[ \rho(s) = s^{-\sigma - 1} \] for some \( \sigma \in (0, 1) \), one obtains a \( \sigma \)-stable CRM. The normalized \( \sigma \)-stable process is defined, in distribution, by the ratio \( \mu/\mu(X) \) (Kingman, 1975). The two–parameter Poisson–Dirichlet (2PPD) process is then defined as follows. Let \( \mathbb{P}_\sigma \) denote the probability distribution of a \( \sigma \)-stable CRM \( \mu \). Then, for any \( \theta > -\sigma \), we can consider a probability measure \( \mathbb{P}_{\sigma,\theta} \) such that
\[
\frac{d\mathbb{P}_{\sigma,\theta}}{d\mathbb{P}_\sigma}(\mu) = \mu^{-\theta}(X).
\]
For any random measure \( \mu_{\sigma,\theta} \) whose probability distribution is \( \mathbb{P}_{\sigma,\theta} \), the random probability measure \( \mu_{\sigma,\theta}/\mu_{\sigma,\theta}(X) \) defines a two–parameter PD process (Pitman and Yor, 1997).

As with the Dirichlet process, a 2PPD process selects, almost surely, discrete probability distributions. An alternative definition represents the 2PPD random probability measure as an infinite mixture, i.e.
\[
\tilde{p} = \sum_{k=1}^{\infty} \tilde{w}_k \delta_{Y_k},
\]
where \( \delta_c \) denotes the point mass at \( c \), the \( \tilde{w}_k \)’s are random weights such that \( \sum_{k=1}^{\infty} \tilde{w}_k = 1 \) (almost surely), and the \( Y_k \)’s are random locations in \( \mathbb{X} \), independent of the weights \( \tilde{w}_k \)’s. If we assume \( \tilde{w}_1 = V_1 \) and \( \tilde{w}_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \), \( k \geq 2 \) with \( V_j \sim \text{Beta}(1 - \sigma, \theta + j\sigma) \), \( j \geq 1 \), then the random probability measure \( \tilde{p} \) coincides in distribution with a 2PPD process (stick-breaking construction; Pitman, 1995, 1996).

Therefore, if the sample \( X_1, \ldots, X_n \) is a realization from an exchangeable sequence driven by a 2PPD process, there is a positive probability of ties, i.e. \( \mathbb{P}[X_i = X_j] > 0 \) for any \( i \neq j \). This clustering property of CRMs often motivates their use in statistical applications, to describe homogeneous subgroups in an heterogeneous population.

In Bayesian hierarchical mixture models, the 2PPD process is often used as an alternative to the mixture of the Dirichlet process (Lo, 1984; Ishwaran and James,
2001), where the specification is completed by assuming that the random locations $Y_k \overset{i.i.d.}{\sim} P^*$, $k \geq 1$, where $P^*$ is an appropriately chosen distribution, which represents the prior expected value of the random distribution, i.e. $E[\tilde{p}(A)] = P^*(A)$ for any $A \in \mathcal{X}$ (base measure). For this reason, in the following we will often succinctly use $X_i | \tilde{p} \overset{iid}{\sim} \tilde{p}$, with $\tilde{p} \overset{d}{=} 2PPD(\sigma, \theta, P^*)$, $i = 1, \ldots, n$ to indicate a sample from a 2PPD with parameters $\sigma$ and $\theta$, and base measure $P^*$.

### 3.2.1 Clustering and diversity

The clustering behavior of the 2PPD process can also be investigated by considering the exchangeable partition probability function (EPPF) which characterizes the distribution of the exchangeable random partition induced by the 2PPD process. A random partition is said to be exchangeable if for any partition $\{A_1, \ldots, A_K\}$ of $n$ elements, the probability of such partition, say $P(\{A_1, \ldots, A_K\})$, is invariant under any permutation of the $n$ elements, and can be expressed as a function of $n = (n_1, n_2, \ldots, n_K)$, where $n_j$ denotes the cardinality of each set $A_j$ for $j = 1, \ldots, K$. Then, the EPPF defines the probability that $X_1, \ldots, X_n$ are partitioned into $K$ distinct clusters with respective sizes $n_1, \ldots, n_K$,

$$
\Pi_K^{(n)}(n_1, \ldots, n_K) = \frac{\prod_{j=1}^{K-1} (\theta + j\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{K} (1 - \sigma)_{n_j-1},
$$

(3.3)

for any choice of positive integers $n_1, \ldots, n_K$ such that $\sum_{i=1}^{K} n_i = n$ and for any number of clusters $K \in \{1, \ldots, n\}$.

Equation (3.3) highlights that the values of the parameters $\sigma$ and $\theta$ affect the clustering structure induced by the two-parameter Poisson–Dirichlet process. As a matter of fact, it is well-known that if $K_n$ denotes the number of distinct values recorded in a sample $X_1, \ldots, X_n$ drawn from an exchangeable sequence directed by a
2PPD(σ, θ) process, then $K_n/n^\sigma \to S_{σ,θ}$ as $n \to ∞$ (almost surely) for some positive random variable $S_{σ,θ}$ (see Theorem 3.8 in Pitman, 2006). Hence, the larger $σ$ is, the larger the number of clusters. Moreover, $σ$ controls the reinforcement of the partition, i.e. the ability of bigger clusters to attract even more observations. This mechanism is evident by considering the predictive distribution of the 2PPD process,

$$
P[X_{n+1} \in A \mid X_1, \ldots, X_n] = \frac{\theta + σK}{θ + n - 1} P^* + \sum_{j=1}^{K} \frac{n_j - σ}{θ + n - 1} \delta_{X_j}(A), \tag{3.4}
$$

where the probability that a new observation is assigned to an existing cluster $X^*$ is proportional to $n_j - σ$. Therefore, values of $σ$ close to 1 favor the formation of a large number of clusters, most of which are singletons (Lijoi et al., 2007).

Alternatively, one can evaluate the clustering structure implied by a Bayesian nonparametric process by computing the expected Simpson’s index of diversity $D_S$. Diversity indexes quantify both the richness (the number of the $p_i$’s) and the evenness (the relative size of the $p_i$’s), and therefore they are often used to characterize the clustering structures in a population. In particular, $D_S$ provides an assessment of the probability that two randomly selected individuals from a population belong to different groups. For a discrete probability distribution $\sum_{i \geq 1} p_i δ_{x_i}$, the Simpson’s index of diversity, is defined as $D_S = 1 - \sum_{i \geq 1} p_i^2$. Then, if $\tilde{p}$ is a 2PPD(σ, θ) process with random probability masses $\{\tilde{p}_j : j \geq 1\}$, then

$$
E[D_S] = \frac{θ + σ}{θ + 1}. \tag{3.5}
$$

Therefore, as $σ$ increases, the population diversity increases, with the highest diversity achieved for $σ = 1$. From equation (3.4), it follows that partitions characterized by high diversity have high probability of containing a large number of singletons and small clusters.
Finally, we consider the variability of realizations from a 2PPD process around its expected value, i.e. the base measure $P^*$, which can be characterized by the variance of the process,

$$\text{Var}[\bar{p}(A)] = \frac{1 - \sigma}{\theta + 1} P^*(A)[1 - P^*(A)]$$

(3.6)

for any $A \in \mathcal{A}$ and $j = 0, 1$. From (3.6), it is evident that large values of $\sigma$ correspond to random probability measures which are more concentrated around the prior guess $P^*$. Therefore, one should expect that the empirical distribution function of any sample $X_1, \ldots, X_n$ drawn from a 2PPD process with high values of $\sigma$, $F_n(b) = \bar{p}(\infty, b) = \sum_{k=1}^\infty \bar{w}_k \delta_{X_k^*}(\infty, b)$, would be characterized by a large number of weights $\bar{w}_k$ of similar size and by $F_n \overset{w}{\rightarrow} P^*$.

### 3.3 A two-groups 2PPD model

The different clustering behavior and diversity that the 2PPD process exhibit as a function of $\sigma$ can be exploited in the two-groups model, to distinguish between the null and alternative distributions. More specifically, we start by rewriting (3.1) as a two-component mixture,

$$\bar{p} = w \bar{p}_0 + (1 - w) \bar{p}_1,$$

(3.7)

where $\bar{p}_0$ and $\bar{p}_1$ represent the unknown distributions under the null and alternative hypothesis, respectively. Here, we specifically assume a $\bar{p}_j \sim \text{2PPD}(\sigma_j, \theta_j, P_j^*)$ process, $j = 0, 1$. Analogous to (3.1), the mixture weight $w$ is a random variable independent of the $\bar{p}_j$'s, and taking values in $[0, 1]$. We further introduce an auxiliary binary random variable $z_i$, $i = 1, \ldots, n$, such that if $z_i = 0$, we assume that $X_i$ is an observation from the null distribution $\bar{p}_0$, whereas if $z_i = 1$, we set $X_i \sim \bar{p}_1$. Thus,
we can rewrite model (3.7) in a Bayesian hierarchical framework as follows,

\[
X_i \mid z_i \overset{\text{iid}}{\sim} \tilde{p}_{z_i}, \quad i = 1, \ldots, n, \\
z_i \mid w \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - w), \\
w \sim \text{Beta}(a, b), \\
\tilde{p}_{z_i} \overset{d}{=} 2\text{PPD}(\sigma_{z_i}, \theta_{z_i}, P^*) .
\]  

(3.8)

Due to the clustering behavior of the 2PPD process, as discussed in Section 3.2.1, we propose to specify the hyper-parameters of the null and non-null random probability measures in (3.8) so that \(\tilde{p}_0\) is concentrated around the theoretical null, in accordance to Efron’s idea that the empirical null distribution should capture only small departures from the theoretical null. On the contrary, we assume that there’s no good model \textit{a priori} for the non-null distribution. Therefore, \(\tilde{p}_1\) is allowed to vary more freely on the space of the alternative distributions. In addition, we expect that for any \(X_i = x \in \mathcal{X}\), the null distribution \(\tilde{p}_0(\infty, x]\), which describes the p-value under the null hypothesis, should not substantially depart from a uniform distribution, i.e. the process should encourage both richness and evenness of the realizations (increased diversity). On the contrary, for the non-null distribution, we should expect a more uneven distribution of the realizations. Thus, based on the considerations above, we propose to set \(\sigma_0 > \sigma_1\). In the next Sections, we will investigate how such a choice might help in discriminating between the null and alternative distribution in the multi-comparison problem.

An underlying assumption of the two-groups model is that if two observations assume the same value, they should not be assigned to different groups. For the 2PPD processes, the following result holds,

\textbf{Lemma 1.} If \(X_1, \ldots, X_n\) are a random sample from an exchangeable sequence \(X\)
governed by a random probability measure as defined in (3.8), with $P_j^*$ being a non-atomic base measure, $j = 0, 1$, then

$$P[X_i = X_j \mid z_i \neq z_j] = 0$$

for $i \neq j$.

The result follows directly from the characterization of the 2PPD process as an infinite mixture, and the fact that $P_j^*$ is non-atomic. A notable consequence of this result is that $X_i$’s belonging to the same cluster are generated by the same PD process, and distinct clusters are generated by distinct random probability measures.

We conclude this Section by discussing the joint partition structure induced by model (3.7) for a sample $X_1, \ldots, X_n \mid \tilde{p} \sim \tilde{p}, n \geq 1$. Let $\Pi_{K,j}^{(n)}(n_1, \ldots, n_K)$ denote the EPPF of process $\tilde{p}_j$, $j = 0, 1$, that is the probability of $n$ observations to be assigned to $K$ different clusters of sizes $(n_1, \ldots, n_K)$. For notational simplicity, we assume that

$$\Pi_{K+1,j}^{(n)}(n_1, \ldots, n_i, 0, n_{i+1}, \ldots, n_{K+1}) \equiv \Pi_{K,j}^{(n)}(n_1, \ldots, n_K),$$

for any $j = 0, 1$ and $n_1, \ldots, n_K \geq 1$ such that $\sum_{i=1}^K n_i = n$. Then, the following result provides the EPPF of the mixture of 2PPD processes, as below:

**Proposition 1.** The EPPF associated to the mixture of 2PPD processes in (3.7) is given by:

$$\Pi_{K}^{(n)}(n_1, \ldots, n_K) = \frac{1}{(a+b)_n} \sum_{i \in \mathcal{I}(0,n_j)} (a)|i|(b)_{n-|i|} \times$$

$$\Pi_{K_0,0}^{(|i|)}(i_1, \ldots, i_k) \Pi_{K_1,1}^{(n-|i|)}(n_1 - i_1, \ldots, n_K - i_K) \quad (3.9)$$

where $i = (i_1, \ldots, i_K)$, $|i| = i_1 + \cdots + i_k$, $K_0 = \text{card}\{j : i_j = n_j\}$ and $K_1 = K - K_0$.

A very similar representation of the EPPF can be deduced by considering mixtures of normalized random measures with independent increments (NRMIs), which are
defined by normalizing completely random measures, i.e. \( \tilde{p}_j = \mu_j / \mu_j(X), \) for \( j = 0, 1. \) See Regazzini et al. (2003), James (2008) and James et al. (2009). In that general case, the EPPF of the mixture can still be represented as in (3.9), where

\[
\Pi^{(n)}_{K,j}(n_1, \ldots, n_K) = \frac{1}{\Gamma(n)} \int_0^\infty u^{n-1} e^{-\psi_j(u)} \prod_{l=1}^K \kappa_{n_l,j}(u) \, du
\]

where \( \kappa_{m,j}(u) := \int_X \alpha(dx) \int_0^\infty s^m e^{-us} \rho(s | x) \, ds, \) and

\[
\psi_j(u) := \int_{X \times \mathbb{R}^+} \left[ 1 - e^{-us} \right] \rho_j(s | x) \, ds \alpha_j(dx).
\]

For example, see Proposition 3 in James et al. (2009).

Direct use of (3.9) is far from trivial. Nonetheless, the expression lends itself to an interesting interpretation: conditional on the assignment of the clusters to either \( \tilde{p}_0 \) or \( \tilde{p}_1, \) the respective random partitions are independent. This is an useful remark for devising a suitable computational algorithm for posterior inference.

As previously discussed, in Bayesian analysis it is often common to assume a hierarchical mixture model with continuous components for a sample \( X_1, \ldots, X_n, \) most notably to overcome the discreteness of the realizations from the 2PDP process, which might be considered inadequate for many applications. Thus, in lieu of (3.7), we can assume \( X_i \mid \tilde{p} \sim \tilde{p}, i = 1, \ldots, n, \) with

\[
\tilde{p} = w \int k(X \mid \theta) \tilde{p}_0(d\theta) + (1 - w) \int k(X \mid \theta) \tilde{p}_1(d\theta), \quad (3.10)
\]

that is each component of the two-groups model is a mixture of 2PDP processes. Here, \( k : X \times \Theta \to \mathbb{R}^+ \) denotes a general kernel such that for any \( \theta \in \Theta \) and some \( \sigma \)-finite measure \( \lambda \) on \( (X, \mathcal{X}) \) one has \( \int_X k(x, \theta) \lambda(dx) = 1. \) For our purposes, it is convenient to set \( X = \mathbb{R} \) and let \( \lambda \) coincide with the Lebesgue measure on \( \mathbb{R} \) so that the previous model defines a prior on the space of density functions on \( \mathbb{R}. \) By
conditioning on the auxiliary group indicator variables \( z_i, i = 1, \ldots, n \), we can rewrite model (3.10) as a hierarchical Bayes nonparametric model as follows:

\[
Y_i \mid \theta_i \overset{\text{ind}}{\sim} k(X_i \mid \theta_i), \quad i = 1, \ldots, n
\]

\[
\theta_i \mid z_i, \tilde{p} \overset{\text{ind}}{\sim} \tilde{p}_{z_i},
\]

\[
z_i \mid w \overset{\text{iid}}{\sim} \text{Bernoulli}(w),
\]

\[
w \sim \text{Beta}(a, b)
\]

\[
\tilde{p}_{z_i} \overset{d}{=} 2\text{PPD}(\sigma_{z_i}, \theta_{z_i}, P^*_{z_i}),
\]

where the \( \theta_i \)'s may indicate either a scalar or a vector parameter. Although we have omitted them in (3.11) for notational simplicity, in addition to the \( \theta_i \)'s, the kernel function \( k(\cdot) \) could depend on additional hyper-parameters, say \( \lambda_i \)'s, which are not relevant for the decision problem, and thus are assigned a separate prior. In the following, we will assume \( k(\cdot \mid \theta_i) = \text{Normal}(\cdot \mid \theta_i) \), and set \( \theta_i = (\mu_i, \tau^2_i) \). To complete the specification of model (3.11), we need to determine the base measures \( P^*_0 \) and \( P^*_1 \).

Following Do et al. (2005), we propose to set the base measures using the Normal-inverse-chi-squared distribution. For \( \tilde{p}_0 \) we set \( \text{NI}_\chi^2(m_0, k_0, \nu_0, s^2_0) \), whereas the base measure for \( \tilde{p}_1 \) can be set as a bimodal mixture distribution \( \frac{1}{2} \text{NI}_\chi^2(-m_1, k_1, \nu_1, s^2_1) + \frac{1}{2} \text{NI}_\chi^2(m_1, k_1, \nu_1, s^2_1) \), for some \( m_1, k_0, k_1, \nu_0, \nu_1, s^2_0, s^2_1 \in \mathbb{R}^+ \) and \( m_0 = 0 \). This choice favors decisions for non-null hypotheses with larger effect sizes since marginally our model could be interpreted as a variant of commonly used spike-and-slab priors in the Bayesian variable selection literature (George and McCulloch, 1993, 1997). The other parameters of the 2PPD process are set such that \( \theta_0 = \theta_1 \) and \( \sigma_0 > \sigma_1 \). In the next Sections, we will investigate the effect of such choices for the multiple comparison problem.
3.4 Posterior inference

Since the posterior distribution of the parameters of interest in model (3.8) or (3.11) is not available in closed form, we employ Markov Chain Monte Carlo techniques for posterior inference. Our primary interest is in the group indicators $z_i$'s, which uniquely identify the random probability measure from which the data $X_i$'s were generated, and, correspondingly, the probability of group membership, $w$. In order to describe posterior inference for the auxiliary variable indicators $z_i$'s, we first clarify the notation for the clustering implied by model (3.8). At any iteration of the MCMC algorithm, the vector of observations $X = (X_1, \ldots, X_n)$ is partitioned into $K$ separate clusters, $K \geq 1$. Let $X_1^*, \ldots, X_K^*$ denote the $K \leq n$ unique values in $X$. We denote the corresponding partition sets by $C_{k,n} = \{i : X_i = X_k^*\}$, $k = 1, \ldots, K$, and by $n_k = |C_{k,n}|$ the cardinality of each set. By virtue of Lemma 1, two observations assigned to the same cluster are also assigned to the same random probability measure. Therefore, let $z_k^*$ be an auxiliary random variable such that $z_k^* = 0$ if the partition set $C_{k,n}$ contains draws from $\tilde{p}_0$, and $z_k^* = 1$ otherwise. Then, for any $i \in C_{k,n}$ one has $z_i = z_k^*$, and the $K$-tuple $z^* = (z_1^*, \ldots, z_K^*) \in \{0,1\}^K$ describes the solution of the multiple testing problem, analogous to the vector $z = (z_1, \ldots, z_n) \in \{0,1\}^n$. Then, posterior samples for $z$ can be immediately derived from the posterior samples of the vector $z^*$ and the configuration of the partition sets $C_{k,n}$, $k = 1, \ldots, K$, which can be obtained by means of a Gibbs sampling scheme.

More specifically, let’s consider the joint probability distribution of the vector $z = (z_1, \ldots, z_n)$, and the partition $C_n = C_{1,n}, \ldots, C_{K,n}$,

\[ \mathcal{L}(z, C_{1,n}, \ldots, C_{K,n}) = \mathcal{L}(z) \mathcal{L}(C_{1,n}, \ldots, C_{K,n} \mid z). \]
The joint distribution of $z$ can be obtained as

$$
\mathcal{L}(z) = \frac{(a)|z|(b)n-|z|}{(a+b)n},
$$

(3.12)

where $|z| = \sum z_i$, and $(x)_n$ denotes the falling factorial, i.e. $(x)_n = x(x-1)(x-2)\cdots(x-n+1)$. By Lemma 1 and Proposition 1, the conditional distribution of $C_{1,n}, \ldots, C_{K,n}$ given $z$ can be written as

$$
\mathcal{L}(C_{1,n}, \ldots, C_{K,n} \mid z) = \prod_{K_{00}}^{(|z|)}(n_1(1-z^*_1), \ldots, n_K(1-z^*_K))
\times \prod_{K_{11}}^{(|z|-1)}(n_z^*_{11}, \ldots, n_Kz^*_K) \prod_{k=1}^{K} \prod_{i \in C_{k,n}} \mathbb{I}(z^*_i)(z_i),
$$

(3.13)

where $K_i = \sum_{k=1}^{K} z^*_k$ and $K_0 = K - K_1$ indicate the number of clusters belonging to $\tilde{p}_1$ and to $\tilde{p}_0$, respectively. Expression (3.12) and (3.13) make evident that it is sufficient to consider only the cluster-based vector of indicators $z^*$ in order to determine the joint probability distribution of the $z$ and the partition $C_n$. Therefore, in order to obtain the full conditional distribution of the $z^*_i$s, we may focus only on the $z^*_k$s and the vector of $K \leq n$ observations in $X^*$. Let $n_{-k} = n - n_k$ denote the number of elements in $X$ after removing those in $C_{k,n}$, and similarly let $z^*_{-k}$ denote the $z^*$ vector with the $k$th entry removed, so that $z^*_{-k} \in \{0,1\}^{K-1}$ Furthermore, for any $i \in C_{k,n}$, $k = 1, \ldots, K$, let $K_{-i,1} = \sum_{l \neq k} z^*_l$ indicate the number of clusters assigned to the non-null distribution $\tilde{p}_1$ after removing cluster $C_{k,n}$. According to the value of $z_i = z^*_k$,

$$
K_{-i,1} = \begin{cases} 
K_1 & \quad z_i = 0, \\
K_1 - 1 & \quad z_i = 1. 
\end{cases}
$$

Based on the previous considerations, we are now in a position to write the full conditional of the $z^*_k$s. For notational simplicity, let $z^*_k = \xi$, where $\xi \in \{0,1\}$. Then,
the full conditional $L_k(\xi \mid z_{-k}^*, X^*, C_n) \propto p_{k, \xi}$ where

$$p_{k, \xi} = (a)_{a-n-k-n\xi} P_{\xi}^* (dX_k^*) \prod_{j \neq k} P_{z_j^*}^* (dX_j^*) \prod_{l \in C_{k,n}} 1{z_l^*}(z_l)$$

$$\times \Pi_{K-K_{-i,1}+\xi,0}^{(n-n-k-n\xi)}(n_1(1-z_1^*), \ldots, n_k(1-\xi), \ldots, n_K(1-z_K^*))$$

$$\times \Pi_{K-K_{-i,1}+\xi,1}^{(n-k+n\xi)}(n_1z_1^*, \ldots, n_k\xi, \ldots, n_Kz_K^*).$$

In particular, the probability $P(z_k^* = 1 \mid z_{-k}^*, X^*, C_n)$ determines the probability that for any $i \in C_{k,n}$, the observations $X_i = X_k^*$ are assigned to the non-null distribution, and can be obtained as

$$L_k(x = 1 \mid z_{-k}^*, X^*, C_n) = \frac{1}{1 + (p_{k,0}/p_{k,1})},$$

where the ratio

$$\frac{p_{k,0}}{p_{k,1}} = \frac{P_{0}^* (dX_k^*)}{P_{1}^* (dX_k^*)} \frac{\Gamma(a + n - n_{-k})}{\Gamma(a + n - n_{-k} - n_k)} \frac{\Gamma(b + n_{-k})}{\Gamma(b + n_{-k} + n_k)}$$

$$\times \Pi_{K-K_{-i,1}+\xi,0}^{(n-n-k-n\xi)}(n_1(1-z_1^*), \ldots, n_k, \ldots, n_K(1-z_K^*))$$

$$\times \Pi_{K-K_{-i,1}+\xi,1}^{(n-k+n\xi)}(n_1z_1^*, \ldots, n_k\xi, \ldots, n_Kz_K^*) \left/ \Pi_{K-K_{-i,1}+\xi,0}^{(n-k)}(n_1z_1^*, \ldots, n_Kz_K^*) \right..$$

The previous expression is valid for all normalized random measures with independent increments. If $\tilde{p}_j$ is a PD($\sigma_j, \theta_j$), $j = 0, 1$ process, then the above ratio further simplifies as:

$$\frac{p_{k,0}}{p_{k,1}} = \frac{P_{0}^* (dX_k^*)}{P_{1}^* (dX_k^*)} \frac{(a + n - n_{-k} - n_k)n_k}{(b + n_{-k})n_k} \frac{(1 - \sigma_0)n_{k-1}}{(1 - \sigma_1)n_{k-1}}$$

$$\times \frac{\theta_0 + (K - K_{-k,1})\sigma_0}{\theta_1 + K_{-k,1}\sigma_1} \frac{(\theta_1 + n_{-k}n_k)}{(\theta_0 + n - n_{-k} - n_k)n_k).}$$

(3.14)

It is worth noting that if the two base measures coincide, i.e. $P_0^* = P_1^*$, then the full conditional does not depend on $X_k^*$; therefore, the probability that $X_k^*$ is a draw from
the non-null depends only on the clustering behavior of the 2PPD process implied by the parameters characterizing \( \tilde{p}_0 \) and \( \tilde{p}_1 \).

Posterior inference on the weight \( w \) in (3.7) is conducted by means of post-MCMC analysis, by approximating the posterior expected value \( E[w \mid \text{data}] \) using the indicators \( \{z^*_t = (z^*_{1,t}, \ldots, z^*_{K,t}) : t = 1, \ldots, T\} \) sampled in the T iterations of the MCMC algorithm. More precisely, if we denote by \( B < T \) the burn-in period of the chain, we can compute the following Monte Carlo approximation of the posterior expected value \( E[w \mid \text{data}] \):

\[
E[w \mid X^*, C_n, z^*] = \frac{a + \sum_{k=1}^{K} n_k (1 - z^*_k)}{a + b + n}.
\]

To summarize, in order to implement a Gibbs sampler for sampling the auxiliary indicators in (3.8), at each iteration \( t = 1, \ldots, T \) we draw each \( z^*_k \), from the full conditional \( \mathcal{L}_k(x = 1 \mid z^*_{-k}, X^*, C_n) \). The vectors \( z^*_t, t = 1, \ldots, T \), can be mapped to the vector \( z \) using the partition sets \( C_n \).

Similar considerations can be repeated for the sampling of the vectors \( z^*_k \) in the hierarchical Bayes model (3.10)–(3.11). Let \( (\theta^*_1, \ldots, \theta^*_K) \) denote the vector of unique values assumed by the \( \theta_i \)'s. Then, because of the conditional independence assumptions in (3.11), the full conditional distribution for each \( z^*_k \) can be obtained

\[
\mathcal{L}[z^*_k \mid z^*_{-k}, \theta^*_1, \ldots, \theta^*_k, C_1, n, \ldots, C_K, n, X] \propto \prod_{\{j: z^*_j = z^*_k\}} P^*(d\theta_j^*) \Gamma(a + \sum_{j \neq k} n_j z^*_j + n_k z^*_k) \Gamma(b + n - \sum_{j \neq k} n_j z^*_j - n_k z^*_k) \\
\times \Pi_{K_k+1, z^*_k} (n_k, n_{1,k}, \ldots, n_{K_k,k}) \mathbb{1}_{\{0,1\}}(z^*_k),
\]
where $K_k = |\{z^*_j = z^*_k : j \neq k\}|$, i.e. the total number of clusters assigned to the same group (hypothesis) as $z^*_k$ in (3.10), $n_{j,k} = |C_{j,n}|$ for any $j \neq k$ such that $z^*_j = z^*_k$, and $n^*_k = n_k + \sum_{r=1}^{K_k} n_{r,k}$, i.e. the total number of observations assigned to the same group as $z^*_k$. Since $w$ is conditionally independent from the observations $X_i$’s given the $z_i$’s and the parameter $\theta_i$, we can obtain the full conditional distribution of $w$ as

$$
\mathcal{L}[dw \mid X_1, \ldots, X_n, \theta_1, \ldots, \theta_n, z_1, \ldots, z_n] = \frac{\Gamma(a + b + n)}{\Gamma(a + n^*)\Gamma(b + n - n^*)} w^{a+n^*-1}(1-w)^{b+n-n^*-1} \mathbb{I}_{(0,1)}(w),
$$

that is a sample from a Beta($a + n^*, b + n - n^*$), with $n^* = \sum_{j=1}^{k} n_j z^*_j$. The sampling algorithm is then completed by drawing samples of the $\theta_i$’s from the respective full conditionals. We detail the full conditionals for the $\theta^*_i$ in the Appendix. Since we assume $k(\cdot \mid \theta_i) = \text{Normal}(\cdot \mid \mu_i, \sigma_i^2)$ and that the base measures $P^*_j$, $j = 0, 1$ are also normally distributed, we can take advantage of the conjugacy properties of the model, and employ a marginal Gibbs sampler as discussed in Ishwaran and James (2001). Alternatively, for non-conjugate models, a Metropolis-Hastings algorithm would be equivalently straightforward to implement, and mimic widely used algorithms for Dirichlet Process mixture models (Neal, 2000).

### 3.5 Identifiability of the two-groups 2PPD model

Identifiability in mixture models is a well-known issue (Teicher, 1963; Titterington et al., 1985; Wasserman, 2000). In a general mixture, $\sum_{j=1}^{J} w_j, f_j(\cdot)$, the mixture components are typically not identifiable, due to the invariance of the mixture under permutation of the indexes. This is usually referred to as the “label switching” problem (Redner and Walker, 1984; Stephens, 2000). In a Bayesian framework, the problem may lead to posterior distributions which are highly multimodal, and therefore
difficult to summarize through commonly used posterior means and marginalization techniques. A common solution has involved some type of relabeling of the parameter space, e.g. such that $w_1 > w_2 > \ldots > w_J$. However, relabeling methods have been shown not to overcome the problem entirely, since the posterior distribution may still retain symmetry and multimodality (see the discussion in Richardson and Green, 1997; Stephens, 2000; Celeux et al., 2000). Celeux et al. (2000) propose to obtain estimators of the parameters of interest in a decision theoretic framework, by designing loss functions for which the lack of labeling is immaterial, e.g. by measuring the distance between two points configurations. The label switching problem is even more accentuated in a Bayesian nonparametric setting, due to the uncertainty related to the estimation of the unknown distributions $f_j(\cdot)$. In a classical setting, Bordes et al. (2006) and Hunter et al. (2007) discuss sufficient conditions for the identifiability of nonparametric mixtures which are symmetric around 0, and with mixture components which can be characterized as location-shifted distributions in a same family, for $K = 2$ and $K = 3$, respectively. In the Bayesian framework, (Kottas and Fellingham, 2010) also focus the two-component symmetric mixture setting, and propose a Bayesian semiparametric model with parametric priors on the unknown location parameters, and the mixture weights, whereas the unknown common $f(\cdot)$ is obtained as a nonparametric scale uniform DP mixture prior, further constrained to be unimodal to facilitate identifiability. Also Mena and Walker (2014) have recently proposed to address the identifiability problem by suitably constructing a prior distribution on the allocation variables $z^*$. More specifically, they propose to employ “asymmetric” priors on the space of the labels, such that the labels are associated with a non-increasing sequence of groups sizes. By enabling the creation of fewer and more interpretable clusters of observations, they show on benchmark datasets that
Table 3.1: Identifiability of the two-groups 2PPD model: Hyper-parameter values for the $\tilde{p}_j$, $j = 0, 1$ used in the illustration discussed in Section 3.5, with the corresponding prior expected number of clusters when $n = 1,000$ and the values of Simpson’s diversity indexes.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>($\sigma_0, \theta_0$)</th>
<th>($\sigma_1, \theta_1$)</th>
<th>$E_{\tilde{p}_0}[K_n]$</th>
<th>$E_{\tilde{p}_1}[K_n]$</th>
<th>$E_{\tilde{p}_0}[D_s]$</th>
<th>$E_{\tilde{p}_1}[D_s]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>(0.90, 1)</td>
<td>(0.10, 1)</td>
<td>578.3</td>
<td>10.9</td>
<td>0.95</td>
<td>0.55</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>(0.75, 1)</td>
<td>(0.25, 1)</td>
<td>256.8</td>
<td>20.8</td>
<td>0.875</td>
<td>0.625</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>(0.67, 1)</td>
<td>(0.33, 1)</td>
<td>167.6</td>
<td>30.1</td>
<td>0.835</td>
<td>0.665</td>
</tr>
</tbody>
</table>

their method effectively improves estimation of the weights and parameters of the mixture components.

Here, we investigate through a numerical example if the different clustering structures implied by different choices of the parameters of the 2PPD process model could facilitate identifiability of the allocation probabilities, hence of the theoretical null distribution, in the two-groups model. We explicitly focus on model (3.7) and, based on the discussion in Section 3.3, we assume $\sigma_0 > \sigma_1$ so that $\tilde{p}_0$ should be more concentrated over the theoretical null distribution. Accordingly, we assume that the base measures of the two processes $\tilde{p}_j$, $j = 0, 1$ are $P^*_0 = \text{Normal}(0, 1)$ and $P^*_1 = \text{Normal}(\mu_1, 1)$, for some $\mu_1 \in \mathbb{R}$. We further assume $\theta_j = 1$, $j = 0, 1$, since our interest is primarily to investigate the effect of the prior constraint.

Here, we present the results of a simulation experiment with $n = 1000$ observations $X_i \sim \tilde{p}$, generated considering 3 different choices of the parameters $\sigma_j$, $j = 0, 1$, which are reported in Table 3.1 together with the corresponding expected value of the Simpson’s diversity index (3.5) and the expected number of clusters generated by
each process (see eq. 3.13 in Pitman, 2006),

\[ E[K_n] = \sum_{i=1}^{n} \frac{(\theta + \sigma)_{i-1}}{(\theta + 1)_{i-1}}. \]

For each scenario, we further consider the following cases: \( \mu_1 = 0 \), corresponding to \( \tilde{p} \) symmetric, and \( \mu_1 = 1.64 \), corresponding to an asymmetric \( \tilde{p} \) where the non-null distribution is centered on the 10% significance cut-off of a two-sided z-test in the classical framework. We also consider two choices for the mixing proportions: \( w = 0.67 \), which correspond to assuming that one third of the observations are drawn from the non-null distribution, as well as the extreme case \( w = 1 \). We fit the data by using model (3.8), where we consider a Beta(1, 2) as a prior on \( w \). Each simulation is repeated 100 times, and for each generated dataset we compute the posterior expected value \( E[w \mid \text{data}] \) as described in Section 3.4.

We summarize the results of our experiment in Figure 3.1, where for each scenario we show the box-plot of the differences between the true \( w \) and \( E[w \mid \text{data}] \). According to expectations, the accuracy of the estimate, hence the identifiability of the mixture components, improves for large \( \mu_i \) and for increasing separation between \( \sigma_0 \) and \( \sigma_1 \). In particular, especially for \( w = 1 \) and \( \mu_i = 0 \), closer values of \( \sigma_0 \) and \( \sigma_1 \) imply more arbitrary assignment of the observations to one of the two processes; however, such effect is considerably reduced for \( \mu_i = 1.64 \). On the other hand, for \( w = 0.67 \), the accuracy of the estimation is improved even for \( \mu_i = 0 \) due to the increased number of observations drawn from the non-null distribution. Overall, these results suggest that appropriate choices of the parameters of the 2PPD processes in the two-groups model can assist, at least partially, in the estimation of the allocation probabilities and the discrimination between the two processes, due to the different implied clustering structures associated with different values of the \( \sigma_j \)'s.
Figure 3.1: Identifiability of the two-groups 2PPD model: Simulation Results under the three Scenarios in Table 3.1, for $\mu_1 \in \{0, 1\}$ and $w \in \{0.67, 1\}$. See Section 3.5 for details.

### 3.6 Simulation study

We investigate the performance of the Bayesian hierarchical 2PPD mixture modeling framework described in (3.10)–(3.11) for large-scale multiple hypothesis testing by means of a simulation study under three scenarios. More specifically, we simulate data from mixture (3.1), where $f_0$ and $f_1$ are density distribution, which represent the distribution of truly null and truly significant $z$-scores, respectively. In our simulations, we set $f_0(z) = N(z \mid \mu, \sigma^2)$, with $\mu = 0$ and $\sigma^2 = 1$, i.e. the theoretical null distribution, and consider the following choices for the non-null distribution:

- **Scenario 1**: $f_1(z) = N(z_i \mid u_i, \sigma^2)$ with $u \sim \text{Uniform}(2, 4)$,
• **Scenario 2**: \( f_1(z) = N(z_i \mid u_i, \sigma^2) \) with \( u \sim \text{Uniform}([-4, -2] \cup [2, 4]) \),

• **Scenario 3**: \( f_1(z) = (-1)^{v_i} \times \text{Gamma}(z_i \mid a, b) \) with \( a = 4, b = 1 \) and \( v_i \sim \text{Bernoulli}(0.5) \).

In these scenarios, z-scores under the non-null hypothesis are drawn from a distribution which is, respectively, asymmetric unimodal (scenario 1), symmetric bimodal (scenario 2), and symmetric bimodal with fat tails (scenario 3), thus mimicking typical high-dimensional testing situation. We consider two values of the mixture proportion \( w \in \{0.90, 0.95\} \), corresponding to the suggestion that in typical large-scale inference hypothesis testing only a small proportion of the observations may be drawn from the non-null distribution. Each simulation includes \( n = 1,000 \) observations, and it is replicated 30 times to allow quantification of posterior uncertainty in the testing problem.

For model fitting, we employ the mixture model (3.10)–(3.11), where we assume \( k(\cdot \mid \theta_i) = \text{Normal}(\cdot \mid \theta_i) \), with \( \theta_i = (\mu_i, \tau_i^2) \). The base measures of the 2PPD process \( \tilde{p}_0 \) is chosen as a Normal-Inverse-chi-squared distribution, i.e. \( P_0^* = \text{NI}\chi^2(m_0, k_0, \nu_0, s_0^2) \), with \( m_0 = 0, \kappa_0 = 1, \nu_0 = 3, s_0^2 = 1 \). For the base measure of \( \tilde{p}_1 \) we consider a mixture \( \frac{1}{2}\text{NI}\chi^2(-m_1, k_1, \nu_1, s_1^2) + \frac{1}{2}\text{NI}\chi^2(m_1, k_1, \nu_1, s_1^2) \), with \( m_1 = 2, \kappa_1 = 1, \nu_1 = 3, s_1^2 = 1 \). The parameters characterizing the clustering behavior of the 2PPD process priors are set to vary in all combinations of \( \sigma_0 \in \{0.5, 0.75\} \), \( \sigma_1 \in \{0.25, 0.5\} \) for each generated dataset, whereas we fix \( \theta_0 = \theta_1 = 1 \). For the Beta prior on \( w \), we set \( a = 1, b = 3 \). For each dataset, the MCMC algorithm was run for 5,000 iterations after a 1,000 iteration burn-in period. The evaluation of posterior convergence was conducted using standard Bayesian convergence diagnostics.
Table 3.2: Simulation study: sensitivity results across different settings for $\sigma_0$ and $\sigma_1$ for the three simulation scenarios considered in Section 3.6 ($w = 0.9$). The values in the table represent the average $F_1$ score, false positive rate (FPR), and true positive rate (TPR) over 30 replicates with corresponding standard deviations.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\sigma_0 = 0.5$</th>
<th>$\sigma_0 = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1 = 0.25$</td>
<td>$\sigma_1 = 0.5$</td>
</tr>
<tr>
<td>F1</td>
<td>0.73 (0.05)</td>
<td>0.74 (0.07)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.06 (0.03)</td>
<td>0.05 (0.03)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.88 (0.04)</td>
<td>0.84 (0.07)</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>F1</td>
<td>0.73 (0.05)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.04 (0.03)</td>
<td>0.01 (0.01)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.78 (0.07)</td>
<td>0.66 (0.06)</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>F1</td>
<td>0.80 (0.03)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.02 (0.01)</td>
<td>0.01 (0)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.79 (0.05)</td>
<td>0.72 (0.06)</td>
</tr>
</tbody>
</table>

We compare the performance of our modeling approach with three alternative methods for large-scale hypothesis testing: (a) the local false discovery rate of Efron (2004); (b) the empirical Bayes mixture model recently proposed by Muralidharan (2010), which allows simultaneous estimation of the effect size and the local false
discovery rate, and which has been shown to outperform an array of alternative methods in simulations from normal data; (c) and the empirical Bayes semiparametric approach of Martin and Tokdar (2012). For each simulation replicate, results were compared using the $F_1$ statistic, which is defined as the harmonic mean of the Precision and Recall statistics, where

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

and

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$ 

The posterior probability that an observation belongs to the non-null group was obtained from the MCMC output as $p(z_i = 1 \mid \text{data}) \approx \frac{1}{T-B} \sum_{t=B+1}^{T} z_{i,t}$, where $z_{i,t}$ denotes the binary indicators sampled for $t = 1, \ldots, T$ iterations of the sampler, with $B$ indicating the burn-in period. For each replicated dataset, we evaluate the performance metrics in our approach by thresholding the posterior probability of the alternative such that $X_i$ is assigned to the non-null distribution if $p(z_i = 1 \mid \text{data}) > 0.5$. For all the other methods, we identify all observations which are deemed significant at a 10% false discovery rate level.

In Table 3.2, we report the performance metrics achieved in the different simulation scenarios, as a function of the combinations of the hyper-parameters of the 2PPD process, for $w = 0.9$. Lower false positive rates are usually associated with the scenarios with a non-null distribution symmetric around zero, as expected since in all cases we consider a symmetric $P^*_1$. Results of the sensitivity analysis for $w = 0.95$ are provided in Table 3.6 in the Appendix and show similar patterns. Overall, the results appear quite robust to the different choices of the hyper-parameters. Table 3.3 reports the results from the comparison with alternative multiple testing methods.
Table 3.3: Simulation study: comparison between the proposed two-groups 2PPD model and three other multiple comparison methods for the three simulation scenarios considered in Section 3.6 ($w = 0.9$). The values in the table represent the average $F_1$ score over 30 replicates with corresponding standard deviations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>0.73 (0.05)</td>
<td>0.59 (0.1)</td>
<td>0.53 (0.11)</td>
<td>0.65 (0.07)</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>0.73 (0.04)</td>
<td>0.53 (0.07)</td>
<td>0.47 (0.09)</td>
<td>0.56 (0.06)</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>0.80 (0.03)</td>
<td>0.69 (0.04)</td>
<td>0.64 (0.06)</td>
<td>0.65 (0.05)</td>
</tr>
</tbody>
</table>

Quite remarkably, the proposed approach outperforms the other methods under each scenario. The increased performance may be attributed to the combined effect of the choice of the hyper-parameters in our model and the lower mass assigned in a neighborhood of the null hypothesis.

### 3.7 Case study: Microbiome data

We illustrate the applicability of the proposed two-groups 2PPD process model for large scale inference on a publicly available gut microbiome dataset, from a case-controlled study of post-diarrheal disruption in children from low-income countries (Pop et al., 2014). Stool samples were obtained from 992 children between the ages of 0 and 59 months, 508 of whom had recently suffered from moderate to severe diarrhea, with the remaining 484 children acting as age-matched controls. The samples were obtained in Mali, The Gambia, Kenya, and Bangladesh and case/control proportions were approximately equal from each country.
Figure 3.2: Microbiome data case study: Histogram of 564 z-scores obtained from the case term ($\beta_1$) in the Negative Binomial generalized linear mixed effects model with point-wise posterior density estimates for the null and non-null distributions.
The purpose of the study was to identify potential microbiota, which may be responsible for exacerbating clinical conditions, and therefore show positive associations with presence of moderate-to-severe diarrhea (MSD) in the case group. Taxa with negative associations are also of interest since they may indicate potential target treatments for recovery from dysbiosis.

Microbial community compositional data were obtained by sequencing highly variable regions of the 16S rRNA gene. Most bacterial species have specific instances of this marker gene; therefore, the 16S rRNA gene can be used to map, at least approximately, individual sequences obtained from a sample into an individual bacterium of a given Operational Taxonomic Unit (OTU, Morgan and Huttenhower, 2012). Thus, the observed sequence abundance serves as a proxy for OTU abundance. In our dataset, the number of OTUs per sample ranged from 55 to 1252, with a median of 380 and an average of 412.

Due to the nature of the sampling mechanism microbiome data are highly skewed, i.e. there are few taxa present in very high abundances and many taxa with low frequencies. Therefore, the Poisson or zero-inflated Poisson, zero-inflated Negative Binomial, as well as other generalized linear regression models have been proposed in the literature to describe the relationship between microbiome data and covariates of interest (Chen and Li, 2016; Shi et al., 2016). Here, we consider a Negative-Binomial regression model on the taxonomic abundances $y_{ij}$, where $j = 1, \ldots, J_i$ indexes the taxa, and $i = 1, \ldots, n$ indexes the observations. As it is typical when dealing with sequencing data (Anders and Huber, 2010; Witten, 2011; Love et al., 2014), we let $s_i$ denote an estimate of a sample-specific size factor, in order to take into account the different sequencing depths of the samples. Also, we let $x_{ij}$ denote a binary covariate, such that $x_{ij} = 1$ if observation $y_{ij}$ is from one of the cases and $x_{ij} = 0$ otherwise.
Then, we consider the following model:

\[
y_{ij} \sim NB(\mu_{ij}, \alpha_j), \quad j = 1, \ldots, J; \quad i = 1, \ldots, n,
\]

\[
\log(\mu_{ij}) = s_i + \beta_{0,j} + \beta_{1,j} x_{ij} + u_{\text{country},j} + u_{\text{case},j},
\]

where \( \alpha_j \) represents a taxon-specific dispersion parameter, \( \beta_{0,j} \) represents a taxon-specific effect, to capture the abundance of the taxon in the control group, and \( u_{\text{country},j} \) and \( u_{\text{case},j} \) are independent normally distributed zero-mean random effects with variances \( \sigma_{\text{country},j}^2 \) and \( \sigma_{\text{case},j}^2 \), in order to capture group-specific variability. The model was chosen due to its flexibility and improved Akaike’s Information Criteria (AIC) scores when compared with Poisson, Zero-inflated Poisson, and Zero-inflated Negative Binomial generalized linear mixed models (Xu et al., 2015). Here, we are interested in the fixed case-control effect from each of the \( \beta_{1,j} \)'s, which we employ to provide the \( z \)-scores for testing the differences in abundance between healthy and MSD subjects. A histogram of the 564 \( z \)-scores is given in Figure 3.2.

We consider the two-groups model (3.10)–(3.11) with hyper-parameters for the prior processes set to \( \theta_0 = \theta_1 = 1 \) and \( \sigma_0 = 0.75 \) and \( \sigma_1 = 0.25 \). The base measure hyper-parameters were set to the same values as in Section 3.6. For the results provided here, we have run the MCMC algorithm described in Section 3.4 for 10,000 iterations after a 1,000 iteration burn-in period. In order to eliminate auto-correlation in the samples, we further thinned the MCMC output to a total of 2,000 samples. No evidence of lack of convergence was observed by running standard diagnostic procedures. Figure 3.2 overlays a point estimate of the posterior densities for the null and non-null densities to the histogram of the \( z \)-scores. Figure 3.3 reports the Monte Carlo estimates of the posterior probability of each taxon belonging to the non-null distribution. By thresholding the Monte Carlo estimate of posterior probability of the
Figure 3.3: Monte Carlo estimates of the posterior probability of each taxon belonging to the non-null distribution.

non-null process at a value corresponding to a Bayesian false discovery rate (Newton et al., 2004) of 5%, we were able to identify a total of 49 non-null taxa, which we report in the tables 3.4 and 3.5. The average number of clusters estimated by the 2PPD process under the null distribution was 47.8 (s.d. 18.9), whereas the average number of clusters estimated under the non-null distribution was 18.2 (s.d. 5.7).

A close inspection of our results reveals some interesting biological findings. Of the eight species which were identified by our method as having significantly less abundance in MDS children, three are *Prevotella* species and three are *Clostridium* species (see Table 3.4). *Prevotella* spp. are common bacteria in the gut and have been found as highly present in children from rural and underdeveloped areas, De Filippo et al. (2010) as well as children whose diets predominantly consist of
carbohydrates and fiber Wu et al. (2011). Thus the severely decreased abundance of the *Prevotella* spp. is reasonable in light of the gastrointestinal disruption MDS children experience. As far as the *Clostridium* spp., *C. difficile*, is well-known as a toxigenic bacteria in adults but it is also found asymptomatically in large proportions in infants and neonates Jangi and Lamont (2010). One of the remaining species is in *Megasphaera* which was recently suggested for reclassification to *Clostridium* (Yutin and Galperin, 2013). The other is *Eubacterium limosum*, one of twenty-five bacterial species recently suggested for fecal transplantation treatment of patients with *C. difficile* infection due to its presence in healthy individuals and resistance to antimicrobial agents Petrof et al. (2013). *Eubacterium limosum* has also been found to be essential for digesting polysaccharides and especially abundant in individuals with high fiber diets Kovatcheva-Datchary et al. (2015).

On the other hand, of the top twenty species (out of the forty-two selected, see Table 3.5) which had been identified as significantly more abundant in the MDS children, ten belong to the *Streptococcus* species. Some *Streptococcus* spp. are well known human pathogens causing conjunctivitis, respiratory infections and urinary tract infections. Other species in the genus are opportunistic pathogens, meaning they are asymptptomatically present in healthy individuals but will flourish in individuals with weakened immune systems such as the patients in this dataset. The pathogenic genus *Shigella* is present and is well-known for causing dysentery. It has been suggested that the *Shigella* spp. are closely related to another well-known pathogen *Escherichia coli* (Lan and Reeves, 2002) which is also differentially abundant in these patients. A *Granulicatella* species has also been identified as differentially abundant. However, this bacteria is usually implicated in childhood infective endocarditis, or infection of the heart. Our findings match many of those discussed in (Pop et al., 2014) (Megas-
phaera, Granulicatella, Streptococcus) while others (Eubacterium limosum, Serratia marcescens) have not been reported in the literature with regard to diarrheal conditions suggesting further investigation of those particular species.

3.8 Discussion and Conclusion

We have considered the two-groups model by Efron (2004) for multiple hypotheses testing, and we have proposed the use of a mixture prior of two-parameter Poisson–Dirichlet processes as a flexible class of prior processes in that framework. In particular, an appropriate choice of the hyper-parameters of the 2PPD processes allows to characterize only small departures from the theoretical null in the estimation of the empirical null distribution, while leaving flexibility in the modeling of the non-null distribution. Such a choice has also been shown to facilitate the estimation of the allocation probabilities and the identifiability of the mixture components. We have also characterized the behavior of the mixture of 2PPD prior by deriving its exchangeable partition probability function, which highlights the conditional independence assumptions in the clustering of the mixture process. The proposed approach has been shown to outperform recently proposed approaches for estimating the components of the two-groups model. Finally, we have illustrated the use of the proposed mixture of 2PPD process on publicly available data from a microbiome study, where we were able to highlight interesting findings previously discussed in the literature, as well as novel ones which will require further validation.

One limitation of our analysis of the microbiome data is related to the fact that many of the species which can now be sequenced using 16S technologies remain still relatively unknown in the microbiology literature, as they cannot be cultured in the
lab. Indeed, it is common for microbiome studies to assign taxa into the genus or even family level so that interpretation is usually conducted at those taxonomic levels. Therefore, the specific function of many taxa is still poorly understood, and interpretation of findings at the taxon level may be difficult.

Another limitation is related to the computing effort, since Markov chain Monte Carlo algorithms for Bayesian nonparametric models typically require considerable computational time for posterior inference. While the data sizes considered in this manuscript are typical of current microbiome studies, a full MCMC approximation of the posterior distributions may become less appealing as sampling techniques improve. For example, metagenomics shotgun sequencing has been increasingly adopted in lieu of 16S rRNA sequencing in human microbiome studies, thus requiring software able to handle the large amount of genomic information being sampled (Sharpton, 2014). Variational Bayes techniques have been developed for many Bayesian nonparametric models, including the 2PPD process (see, e.g. Blei and Jordan, 2006; Sudderth and Jordan, 2009; Broderick et al., 2013a). Fast maximum a posteriori estimates which include structural variables explicitly in the posterior distribution using combinatorial functions such as the EPPF have also been recently proposed in the literature (Broderick et al., 2013b).

Finally, we have proposed a two-groups model for the analysis of differential abundance in taxa observed under two conditions. However, often the interest is in studying how taxa abundance varies longitudinally over time even within a single patient. Therefore, models that take into account the temporal dependence of the hypotheses over multiple taxa may be required. These are all avenues we plan to address in future research.
3.9 Appendix

3.9.1 Full conditional

The full conditional of \( \theta_i = (\mu_i, \tau_i^2) \) depends on the predictive distributions of \( \tilde{p}_1 \) and \( \tilde{p}_0 \) and will coincide with

\[
\mathcal{L}[d\theta_i \mid \theta_{-i}, z^*, C_n, X] \propto q_0^{(i)} \int_{\mathbb{R} \times \mathbb{R}^+} k(X_i \mid \theta_i) \alpha(d\theta) + \sum_{j=1}^{K_{z_i}} q_j^{(i)} \delta_{\theta_j} \alpha(d\theta)
\]

where \( \theta_{-i} = \{ (\mu_j, \sigma_j^2) : j \neq i \} \) and \( k(X \mid \theta) \) is the Gaussian density function and \( K_{z_i} \) is the number of unique values in \( \theta_{-i} \) that share the same generating random probability measure \( \tilde{p}_{z_i} \) with \( \theta_i \). Correspondingly, the respective frequencies are denoted as \( n_{z_i} = n_{1,z_i}, \ldots, n_{K_{z_i},z_i} \) with weights \( q_0^{(i)} \) and \( q_j^{(i)} \) as follows

\[
q_0^{(i)} \propto \prod_{K_{z_i}+1}^{[n_{z_i}+1]} (n_{1,z_i}, \ldots, n_{K_{z_i},z_i}, 1) k(X_i \mid \theta_i) \\
q_j^{(i)} \propto \prod_{K_{z_i}}^{[n_{z_i}+1]} (n_{1,z_i}, \ldots, n_{j,z_i}+1, \ldots, n_{K_{z_i},z_i})
\]

which must clearly be such that

\[
q_0^{(i)} + \sum_{j=1}^{K_{z_i}} q_j^{(i)} = 1.
\]

3.9.2 Proof of Proposition 1.

We first evaluate the probability distribution of the \( X_i \)'s, partitioned into \( K \) distinct clusters with representatives located at infinitesimal intervals \( dx_1, \ldots, dx_K \), around points \( x_1, \ldots, x_K \), with respective multiplicities \( n_1, n_2, \ldots, n_K \).

\[
P[X_1^* \in dx_1, \ldots, X_K^* \in dx_K, n_1, \ldots, n_K] = E \left[ \prod_{j=1}^{K} \left\{ \frac{w \mu_0(dx_j)}{\mu_0(X)} + (1-w) \frac{\mu_1(dx_j)}{\mu_1(X)} \right\}^{n_j} \right]
\]
\[
= \sum_{i_1=0}^{n_1} \cdots \sum_{i_K=0}^{n_K} \binom{n_1}{i_1} \cdots \binom{n_K}{i_K} w^{i_1+\cdots+i_K} (1-w)^{n-(i_1+\cdots+i_K)} \times
\]
\[
E \left[ \prod_{j=1}^{K} \left( \frac{\mu_0(dx_j)}{\mu_0(\mathbb{X})} \right)^{i_j} \left( \frac{\mu_1(dx_j)}{\mu_1(\mathbb{X})} \right)^{n_j-i_j} \right]
\]

From Proposition 2, it follows that for any \( i_j \notin \{0, n_j\} \) the expected value above vanishes. Hence, for \( i_j \in \{0, n_j\} \) one has

\[
E \left[ \prod_{j=1}^{K} \left( \frac{\mu_0(dx_j)}{\mu_0(\mathbb{X})} \right)^{i_j} \left( \frac{\mu_1(dx_j)}{\mu_1(\mathbb{X})} \right)^{n_j-i_j} \right] = \prod_{j=1}^{K} [P_0^*(dx_j)]^{i_j} [P_1^*(dx_j)]^{n_j-i_j}
\]

\[
= \prod_{|i|,0}^{(n'_{i'})} (i_1, \ldots, i_K) \prod_{|i|,1}^{(n-n'_{i'})} (n_1 - i_1, \ldots, n_K - i_K)
\]

The representation in Proposition 3, then, follows when integrating out with respect to \( w \) and \( x_1, \ldots, x_k \).

\[\square\]

3.9.3 Sensitivity results
Table 3.4: Microbiome data case study: differentially abundant taxa with negative z-scores indicating less abundance in the children with moderate to severe diarrhea. Most are well known commensal bacteria, e.g. *Prevotella* spp. and *Clostridium* spp.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>z-score</th>
<th>MPPI</th>
<th>local FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megasphaera sp. TrE9262</td>
<td>-7.82</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Clostridium paraputrificum</td>
<td>-7.50</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Prevotella sp. DJF_B112</td>
<td>-6.73</td>
<td>0.99</td>
<td>0.88</td>
</tr>
<tr>
<td>Clostridium sp. FRC_Cl1</td>
<td>-6.17</td>
<td>0.98</td>
<td>0.76</td>
</tr>
<tr>
<td>Prevotella sp. BI-42</td>
<td>-6.11</td>
<td>0.98</td>
<td>0.74</td>
</tr>
<tr>
<td>Prevotella copri</td>
<td>-6.05</td>
<td>0.99</td>
<td>0.72</td>
</tr>
<tr>
<td>Clostridium sp. CJ66</td>
<td>-5.83</td>
<td>0.99</td>
<td>0.64</td>
</tr>
<tr>
<td>Eubacterium limosum</td>
<td>-5.41</td>
<td>0.98</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Table 3.5: Microbiome data case study: differentially abundant taxa with positive $z$-scores indicating greater abundance in the children with moderate to severe diarrhea. Most are well known pathogenic bacteria, e.g. *Shigella* spp. and *E. coli*.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>$z$-score</th>
<th>local FDR</th>
<th>local FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Streptococcus mitis</em></td>
<td>12.33</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral clone ASCE09</em></td>
<td>10.64</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>10.55</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><em>Shigella sp. DBC-1</em></td>
<td>7.68</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td><em>Granulicatella elegans</em></td>
<td>7.47</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em></td>
<td>7.40</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral clone DP009</em></td>
<td>7.39</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral clone ASC01</em></td>
<td>7.29</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td><em>Neisseria sp. oral clone BP2-82</em></td>
<td>7.25</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td><em>Shigella boydii</em></td>
<td>6.69</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td><em>Granulicatella sp. oral clone ASC04</em></td>
<td>6.59</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral clone BP1-49</em></td>
<td>6.50</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral clone ASC04</em></td>
<td>6.40</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td><em>Streptococcus sp. C101</em></td>
<td>6.29</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td><em>Enterobacter cloacae</em></td>
<td>6.25</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td><em>Streptococcus sp. oral strain T4-E3</em></td>
<td>6.21</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td><em>Streptococcus pasteurianus</em></td>
<td>6.13</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td><em>Serratia marcescens</em></td>
<td>5.97</td>
<td>0.99</td>
<td>0.87</td>
</tr>
<tr>
<td><em>Streptococcus sanguinis</em></td>
<td>5.97</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td><em>Lactobacillus fermentum</em></td>
<td>5.80</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td><em>Streptococcus peroris</em></td>
<td>5.74</td>
<td>0.99</td>
<td>0.84</td>
</tr>
<tr>
<td><em>Haemophilus sp. oral clone BJ021</em></td>
<td>5.72</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td><em>Aggregatibacter segnis</em></td>
<td>5.66</td>
<td>0.99</td>
<td>0.82</td>
</tr>
<tr>
<td><em>Erwinia sp. oral clone</em></td>
<td>5.64</td>
<td>0.99</td>
<td>0.82</td>
</tr>
<tr>
<td><em>Actinobacillus pleuropneumoniae</em></td>
<td>5.42</td>
<td>0.99</td>
<td>0.77</td>
</tr>
<tr>
<td><em>Actinomyces odontolyticus</em></td>
<td>5.31</td>
<td>0.99</td>
<td>0.76</td>
</tr>
<tr>
<td><em>Weissella cibaria</em></td>
<td>5.20</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td><em>Citrobacter freundii</em></td>
<td>5.14</td>
<td>0.99</td>
<td>0.71</td>
</tr>
<tr>
<td><em>Haemophilus haemolyticus</em></td>
<td>4.96</td>
<td>0.99</td>
<td>0.66</td>
</tr>
<tr>
<td><em>Unresolved Taxon</em></td>
<td>4.85</td>
<td>0.99</td>
<td>0.63</td>
</tr>
<tr>
<td><em>Fusobacterium sp. oral clone DJF</em></td>
<td>4.68</td>
<td>0.99</td>
<td>0.57</td>
</tr>
<tr>
<td><em>Escherichia albertii</em></td>
<td>4.41</td>
<td>0.99</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Table 3.6: Simulation study: sensitivity results across different settings for $\sigma_0$ and $\sigma_1$ for the three simulation scenarios considered in Section 3.6 ($w = 0.95$). The values in the table represent the average $F_1$ score, false positive rate (FPR), and true positive rate (TPR) over 30 replicates with corresponding standard deviations.

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>$\sigma_0 = 0.5$</th>
<th>$\sigma_0 = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1 = 0.25$</td>
<td>$\sigma_1 = 0.25$</td>
</tr>
<tr>
<td>$\sigma_0 = 0.5$</td>
<td>0.67 (0.1)</td>
<td>0.74 (0.05)</td>
</tr>
<tr>
<td>$\sigma_0 = 0.75$</td>
<td>0.66 (0.09)</td>
<td>0.67 (0.1)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.04 (0.02)</td>
<td>0.02 (0.01)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.82 (0.08)</td>
<td>0.78 (0.08)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>$\sigma_0 = 0.5$</th>
<th>$\sigma_0 = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1 = 0.25$</td>
<td>$\sigma_1 = 0.25$</td>
</tr>
<tr>
<td>$\sigma_0 = 0.5$</td>
<td>0.66 (0.09)</td>
<td>0.67 (0.1)</td>
</tr>
<tr>
<td>$\sigma_0 = 0.75$</td>
<td>0.3 (0.02)</td>
<td>0.01 (0.02)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.03 (0.02)</td>
<td>0.01 (0.02)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.73 (0.08)</td>
<td>0.61 (0.09)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 3</th>
<th>$\sigma_0 = 0.5$</th>
<th>$\sigma_0 = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_1 = 0.25$</td>
<td>$\sigma_1 = 0.25$</td>
</tr>
<tr>
<td>$\sigma_0 = 0.5$</td>
<td>0.74 (0.08)</td>
<td>0.75 (0.06)</td>
</tr>
<tr>
<td>$\sigma_0 = 0.75$</td>
<td>0.3 (0.02)</td>
<td>0.00 (0)</td>
</tr>
<tr>
<td>FPR</td>
<td>0.02 (0.02)</td>
<td>0.00 (0)</td>
</tr>
<tr>
<td>TPR</td>
<td>0.75 (0.08)</td>
<td>0.65 (0.07)</td>
</tr>
</tbody>
</table>
Chapter 4

Conclusions and discussion

In this thesis we have presented two approaches to modeling microbiome data. The first explicitly models abundances and the associations between taxa and a related dataset. This modeling approach allows researchers to narrow their search for taxa and their associated covariates such as demographics, metabolomics, or host gene expression by applying state-of-the-art Bayesian variable selection techniques. Efficient exploration of the high-dimensional parameter space was achieved through the use of adaptive Metropolis-Hastings proposal mechanisms. Demonstration of the approach on publicly available data gave interpretable results in congruence with the prevailing literature. This model serves as a useful starting point for several other approaches which could have useful interpretation in microbiome studies. The microbiome is a dynamic system and longitudinal effects could be incorporated into the model we have proposed. If abundance measurements are taken over time we may, for example, make additional assumptions about the intercept term in the Dirichlet-Multinomial regression model and explicitly model the autocorrelation between the longitudinal samples. Another extension involves applying a distinct variable selection mechanism to partitions of a sample. Such an extension would employ a nonparametric process prior which replaces the finite dimensional spike-and-slab parametric prior used previously. These partitions may or may not be known a priori to the practitioner. Heterogeneity in genomics data is common due to cluster sampling and batch effects – among other sources – and such a model would help account for those effects.
The second project uses a more general Bayesian nonparametric framework to perform multiplicity correction for differential abundance studies. We proposed a novel mixture of two-parameter Poisson-Dirichlet processes which may be used as a prior for the two-groups model. By fixing appropriate hyper-parameters to the process we achieve improved differentiation between the observations in the null and non-null groups. Straight-forward computation of the posterior is implemented through Gibbs sampling in the conjugate case. This approach could be even better tailored to microbiome data by incorporating a more sophisticated method for obtaining the difference scores. For example, in the style of Zhao et al. (2014) the upper-level mixture could be informed, not by external prior information, but by interactions between the taxa.

Throughout we have explicitly addressed the count-based aspect of microbiome abundance tables. We have advocated for specific probability distributions which respect the structure of the data. However, other analysis approaches have received attention recently (Friedman and Alm, 2012; Mandal et al., 2015; Shi et al., 2016; Cao et al., 2016) and may be worthwhile to explore in a Bayesian context. Compositional data analysis (Aitchison, 2003) proposes a method for transforming data that exists on the constrained unit simplex into the more familiar real space. After such transformations have been made it is straightforward to apply standard multivariate statistical procedures. To make these statements more concrete, let us consider that a microbiome sample consists of positive (and zero) entries of absolute abundance. That is, a microbiome sample \( \mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}) \in \mathbb{N}^J \) is a vector of counts of the number of taxa detected in the sample. A simple normalization \( \mathbb{R}^J \rightarrow S^{J-1} \) to the unit simplex \( \mathbf{y}_i^* = (y_{i1}/\sum_j y_{ij}, y_{i2}/\sum_j y_{ij}, \ldots, y_{ij}/\sum_j y_{ij}) \in S^{J-1} \) transforms the entries of \( \mathbf{y}_i^* \) into relative abundances since the vector \( \mathbf{y}_i^* \) now sums to one. The vector
\( y_i^* \) is now what is known as “compositional data.” However, note that the elements of \( y_i^* \) are now conditionally dependent since we have \( y_{ij}^* = 1 - \sum_{k \neq j} y_{ik}^* \). Compositional data analyses uses, for example, the \textit{additive log ratio}

\[
\mathbf{y}_i^\dagger = (\log(y_{i,1}^*/y_{i,J}), \log(y_{i,2}^*/y_{i,J}), \ldots, \log(y_{i,J-1}^*/y_{i,J})) \in \mathbb{R}^{J-1}
\]

which takes \( S^J \rightarrow \mathbb{R}^{J-1} \) and frees the elements of \( \mathbf{y}_i^\dagger \) from the simplex constraint. However, such a transformation can complicate interpretation since all values become relative to the normalizing quantity \( y_{i,J}^* \).

Notably, classical compositional data analysis cannot handle zeros due to the logarithmic transformation though several authors have considered workarounds (Martin-Fernandez et al., 2000; Martín-Fernández et al., 2003). Furthermore, it is not clear which taxon should act as the normalizing quantity \( y_{i,J}^* \) in such transformations though in applications a particular taxon which is of little interest could be chosen.

The microbiome is a dynamic system with many taxa interacting on a temporal scale (Caporaso et al., 2011; Gerber et al., 2012). Furthermore, taxa compete for space and resources not just with each other, but with their environment. Ecologists have long thought about species interactions and have enumerated all the possibilities (Table 4.1) and there is enormous interest in understanding these interaction in the microbiome (Faust and Raes, 2012). A better understanding of these interactions on a taxon by taxon level will be tremendously difficult due to the dimensionality but may be tremendously rewarding. The ecological concept of the “keystone” species (Mills et al., 1993), where a single species has an outsized effect on the environment around it, including modulating abundances of other species, has been proposed to exist in metagenomic communities (Steele et al., 2011). Future efforts using longitudinal Dirichlet-Multinomial or Negative Binomial models with explicit interaction terms.
Effect on host | Effect on guest | Name
---|---|---
+ | + | Mutualism (symbiosis)
- | - | Competition
0 | 0 | Neutralism
0 | - | Amensalism
+ | 0 | Commensalism
+ | - | Antagonism

Table 4.1: List of possible interactions between microbiome and host. Note that each taxon will interact with every other taxon in one of these ways too.

may be one way to approach this.

As the field of microbiome research continues to grow there will be more need for dedicated statistical tools. Here we have highlighted several other approaches we plan to explore in the future. Further research into compositional data analysis, clustered variable selection models, and models for longitudinal interaction should yield fruitful results in both the statistics and biomedical literatures. Hopefully this thesis work will serve as a starting point for such research in the future.
Bibliography


produced by commensal bacteria promote peripheral regulatory T-cell generation. 


Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y.,
Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L.,
Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen,
T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E.,
Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li,
S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K.,
and Wang, J. (2012). A metagenome-wide association study of gut microbiota in


normalized random measures with independent increments. *The Annals of Statis-

Rho, J. H., Wright, D. P., Christie, D. L., Clinch, K., Furneaux, R. H., and Roberton,
A. M. (2005). A novel mechanism for desulfation of mucin: Identification and
cloning of a mucin-desulfating glycosidase (sulfoglycosidase) from Prevotella strain

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with
an Unknown Number of Components. *Journal of the Royal Statistical Society, B


Wasserman, L. A. (2000). Asymptotic inference for mixture models by using data-


